JMIR Medical Informatics

Impact Factor (2022): 3.2
Volume 11 (2023) ISSN 2291-9694 Editor in Chief: Christian Lovis, MD, MPH, FACMI

Contents

Reviews

Original Papers

Viewpoints

Implementation Reports

Corrigenda and Addendas

Editorials

Machine Learning–Enabled Clinical Information Systems Using Fast Healthcare Interoperability Resources Data Standards: Scoping Review

Jeremy A Balch[1,2,*], MD; Matthew M Ruppert[2,3,*], MA; Tyler J Loftus[1,2], MD; Ziyuan Guan[2,3], MA; Yuanfang Ren[2,3], PhD; Gilbert R Upchurch[1], MD; Tezcan Ozrazgat-Baslanti[2,3], PhD; Parisa Rashidi[2,4], PhD; Azra Bihorac[2,3], MSc, MD

1

2

3

4

*these authors contributed equally

Corresponding Author:
Azra Bihorac, MSc, MD

Abstract

Background: Machine learning–enabled clinical information systems (ML-CISs) have the potential to drive health care delivery and research. The Fast Healthcare Interoperability Resources (FHIR) data standard has been increasingly applied in developing these systems. However, methods for applying FHIR to ML-CISs are variable.

Objective: This study evaluates and compares the functionalities, strengths, and weaknesses of existing systems and proposes guidelines for optimizing future work with ML-CISs.

Methods: Embase, PubMed, and Web of Science were searched for articles describing machine learning systems that were used for clinical data analytics or decision support in compliance with FHIR standards. Information regarding each system's functionality, data sources, formats, security, performance, resource requirements, scalability, strengths, and limitations was compared across systems.

Results: A total of 39 articles describing FHIR-based ML-CISs were divided into the following three categories according to their primary focus: clinical decision support systems (n=18), data management and analytic platforms (n=10), or auxiliary modules and application programming interfaces (n=11). Model strengths included novel use of cloud systems, Bayesian networks, visualization strategies, and techniques for translating unstructured or free-text data to FHIR frameworks. Many intelligent systems lacked electronic health record interoperability and externally validated evidence of clinical efficacy.

Conclusions: Shortcomings in current ML-CISs can be addressed by incorporating modular and interoperable data management, analytic platforms, secure interinstitutional data exchange, and application programming interfaces with adequate scalability to support both real-time and prospective clinical applications that use electronic health record platforms with diverse implementations.

(JMIR Med Inform 2023;11:e48297) doi:10.2196/48297

KEYWORDS

ontologies; clinical decision support system; Fast Healthcare Interoperability Resources; FHIR; machine learning; ontology; interoperability; interoperable; decision support; information systems; review methodology; review methods; scoping review; clinical informatics

Introduction

Data analytic tools provide essential contributions to scientific investigation and clinical decision-making [1]. These tools are in turn fueled by the volumes of data that have been generated since the passage of the Health Information Technology for Economic and Clinical Health Act in 2009, which incentivized the adoption of electronic health record (EHR) systems [2-4]. EHR data, however, remain nonstandardized across institutions and, within an institution, may not be readily available for real-time analysis, thus impairing multi-institutional research efforts and care for individual patients across institutions [5-8]. The standards herein refer to the structure, organization, representation, and transmission of data. Health information exchange systems can mitigate these issues by using the Fast Healthcare Interoperability Resources (FHIR; pronounced "fire") data standard [9]. The Health Level 7 (HL7) International standard developing organization sought to reduce the

complexity of the HL7 version 3 Reference Information Model while maintaining semantic interoperability and thus adopted the FHIR standard in 2011 [10]. It supports multiple development platforms and has been embraced by major industry and government organizations. Since 2016, developers have engaged with Substitutable Medical Applications and Reusable Technologies (SMART) on FHIR to build EHR and commercial applications [11,12]. Despite the growth of technologies using FHIR standards, there is limited literature summarizing differences among machine learning–enabled clinical information systems (ML-CISs), and the best methods for applying FHIR remain unclear.

This review describes the functionalities, strengths, and weaknesses of clinical applications that use the FHIR standard and have been described in the medical literature, and we propose guidelines for improved multi-institutional research initiatives and clinical applicability.

Methods

Given the rapidly evolving nature of this field, we performed a scoping review to provide a critical appraisal of the current literature, with the goal of informing future studies. We followed the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines; the PRISMA-ScR checklist is available in Multimedia Appendix 1.

Research Protocol

We sought articles describing clinical decision support (CDS) systems (CDSSs) or risk prediction systems using FHIR standards. FHIR standards define resource types (ie, patients, medications, and clinical observations), data elements (ie, medication name and dosage), data formats (ie, JSON and XML files), and the use of standard ontologies (ie, Systematized Nomenclature of Medicine-Clinical Terms [SNOMED CT] and Logical Observation Identifiers Names and Codes [LOINC]), among others. Our initial search was performed on April 23, 2020, and given the progress of the field, it was updated again on October 11, 2022. Inclusion criteria involved all full-text articles published in English. We excluded abstracts, poster presentations, and meeting summaries. Embase, PubMed, and Web of Science were searched for cohort studies, case-control studies, and reviews. Our search terms for each database are found in Multimedia Appendix 2. Despite their increasing use by commercial entities, we did not search for commercial applications of FHIR, as their lack of peer review and limited reportability prevented a formal evaluation of their methods. Following the removal of duplicates, 153 articles were identified. Titles and abstracts were reviewed by 2 authors independently, with disagreements resolved by a third. Full-text articles that did not adequately describe system functionality, data sources, formats, security, performance, resource requirements, scalability, strengths, and limitations were excluded. We also excluded articles that described a model architecture using FHIR but did not incorporate it into a CDSS. A total of 39 full-text articles were included for full analysis.

Article Evaluation

Strengths and limitations of the applications were evaluated in terms of functionality, data sources, formats, security, performance, resource requirements, and scalability. *Functionality* was defined as the intended purpose of the algorithm and its capabilities, ranging from the integration of genomic data into the EHR [13,14] to CDSSs [15-17] and predictive models [18,19]. Data sources included information within electronic health care records and external sources, such as wearable devices [20]. Formats were evaluated based on system architecture and the technologies underlying the algorithms (eg, use of Bayesian networks [16], transformers [21], or rule-based methods [22]). Security was evaluated based on how the application handled sensitive health information, including encryption [23], use-and-access control mechanisms [24], or authorization platforms [25,26]. Performance and resource requirements refer to the processing time, memory, and computing needs of the applications. Finally, scalability refers to the likelihood of adoption by other health care systems or platforms (eg, use of open-source components [27] or cloud-based repositories [28]). Knowledge from the included articles was used to propose avenues of future development for optimizing machine learning–enabled systems.

Results

A total of 39 clinical tools that used FHIR standards were divided into the following three categories according to their primary focus: CDSSs (n=18), interoperable data management and analytic platforms (n=10), or auxiliary modules and application programming interfaces (APIs; n=11) that enhance ML-CISs.

The CDSSs

CDSSs are algorithms that use health information to provide assistance for clinical decision-making tasks. Table 1 shows articles that focused on these support systems. Although many CDSSs lacked interoperability and external validity, several characteristics of CDSSs harbored potential for improving both efficacy and efficiency.

Table . Summary of intelligent clinical decision support systems.

Source, year	Functionalities	Strengths	Limitations
Curran et al [15], 2020	Summarizes chronic obstructive pulmonary disease information, provides decision support, and suggests orders	Dynamic embedding within EHR[a] and compatible with SMART[b]-on-FHIR[c] submodules	Limited generalizability due to single center and single disease
Dolin et al [13], 2018	Uses drug-gene interaction data for clinical decision support triggered by EHR medication orders	Accesses a rules engine containing level A recommendations from a pharmacogenetics consortium	Difficult to query the rules engine for level A recommendations when triggered by EHR medication orders
El-Sappagh et al [28], 2019	Uses mobile health technologies to monitor and manage type 1 diabetes	Most system processes are executed in the cloud; once configured, it runs on any EHR system	The diabetes treatment ontology did not address emergency conditions and was not embedded within an EHR system
Gaebel et al [16], 2016	Generates digital patient models for clinical decision support for laryngeal cancer	Bayesian networks are well-suited for representing complex diseases	System architecture was described, but the system was not implemented clinically
Gruender et al [14], 2019	Combines next-generation sequencing genomics data with FHIR clinical data	Open-source system that combines data formats and is portable	Manual data extraction and web-based filtering tool
Gordon et al [29], 2017	Displays patients' thrombocytopenia trends along with computer-generated calculated panel reactive antibody levels	Provides real-time services and effective visual cues	Data sources are limited
Henry et al [18], 2018	Predicts sepsis among intensive care unit patients in real time	Cloud-based system that provides alerts to clinicians	Public cloud-based solutions present safety issues
Hong et al [27], 2019	Phenotypes diabetes based on free-text notes and other structured data	Converts unstructured, semistructured, and structured data to appropriate FHIR components	Performance is not stable across different data sets
Kawamoto et al [30], 2021	Takes data from multiple EHRs and incorporates them into existing risk calculators	Performance measured with end user satisfaction studies and used existing application programming interface	Tested at a single institution and had data security concerns
Park et al [31], 2022	Personal health record application for employees, with links to health care resources	FHIR-based cloud application that is applicable to multiple EHRs and provides secure access through Azure	Limited integration of hospital data
Schleyer et al [32], 2021	Integrates selected data from statewide data systems into local EHR	Translates data from diverse sources into a common database	Experience limited to a single EHR
Semenov and Kopanitsa [20], 2018	Recommends clinical decisions and actions based on EHR data	Free-text output for both physicians and patients	No standard performance evaluation
Semenov et al [33], 2018	Recommends clinical decisions and actions based on EHR data	Free-text output for both physicians and patients and improved analytic workflow relative to prior versions	No standard performance evaluation
Séroussi et al [34], 2018	Produces clinical practice guideline services for patients with breast cancer	Uses both data models and knowledge models and provides effective data analytic visualizations	Implemented on a small scale, proposed guidelines were not validated, and interguideline conflicts need to be resolved manually
Tarumi et al [35], 2021	Modeling of treatment outcomes for type 2 diabetes	Effective use of SMART on FHIR for integration in local EHR, and design incorporated clinician feedback	No external validation, limited access to cost data, and not yet compatible with all EHRs
Thayer et al [36], 2021	Automated graphical display of asthma history	Smoothly integrated into EHR	Not based on SMART, limiting interoperability
Wang et al [37], 2019	Comparison of machine learning algorithms for prediction of end-stage renal disease in type 2 diabetes	Extraction of EHR data using FHIR	Single institution, no imputation of missing data, and no external validation

Source, year	Functionalities	Strengths	Limitations
Whitaker et al [38], 2022	Machine learning algorithm to identify blood transfusion adverse events	Synthesized structured and unstructured data from EHR to achieve reasonable accuracy compared to clinicians	Retrospective study more aligned toward research than clinical care

[a]EHR: electronic health record.

[b]SMART: Substitutable Medical Applications and Reusable Technologies.

[c]FHIR: Fast Healthcare Interoperability Resources.

CDSS ontologies are a central tenant of CDSS interoperability. Generally, ontologies are a hierarchy of concepts that are defined by both a set of attributes and their relationships to other concepts, and they must meet several internal consistency and version control objectives [10]. Common ontologies include the SNOMED CT, LOINC, and National Cancer Institute Thesaurus (NCIT). Separate ontologies may conflict, such as in cases where models use different organizing principles, have varying degrees of granularity, or even exhibit contextual differences between clinical applications and biomedical research. Séroussi et al [34] faced this problem when creating a guideline for the optimal management of breast cancer by integrating a collection of pre-existing ontologies (NCIT and LOINC). They were able to resolve this conflict by using data visualization techniques and rules-based inference engines, though often their methods required the manual resolution of conflicts. Common ontologies can also omit essential elements. Dolin et al [13] were able to transform a library of drug-gene interactions into an FHIR standard to alert physicians when prescriptions are likely to cause adverse drug reactions. Specific disease classes may lack an interoperable ontology. For cancer, there are active efforts in the CodeX HL7 FHIR Accelerator community to capture oncologic data from the EHR by using the mCODE (minimal Common Oncology Data Elements) ontology [39,40].

Advanced CDSSs have been integrated with machine learning algorithms to process data, especially unstructured data, such as clinician notes. Gaebel et al [16] created a physician-facing CDSS that used Bayesian networks and medical language modules to identify the optimal management strategy for laryngeal cancer. Bayesian networks and other modeling approaches can estimate and infer unobserved but relevant variables, which is advantageous in representing complex diseases. Natural language processing is becoming an increasingly common tool. Hong et al [27], Semenov et al [20,33], and Whitaker et al [38] used semantic tags, rules-based extraction, and the scispaCy-based natural language processing pipeline to extract their concepts, though these methods require arduous labeling—the process of manually highlighting terms and classifying them—and lack validation on external data sets. Vocabulary and expressions often differ outside of the training context, requiring developers to further refine their language models after release by using test data and real-life examples.

Cloud-based solutions have made it possible to process large-scale and heterogeneous data and push the boundaries of CDSSs to encompass broader scenarios. El-Sappagh et al [28] developed a mobile app that integrates data from wearable monitors (eg, vital signs, physical activity, and blood glucose levels) with the EHR to provide recommendations for managing type 1 diabetes mellitus. The system delivers spoken education and lifestyle recommendations to patients' mobile devices, using an ontology generated from clinical practice guidelines, expert opinions, and other published sources. Meanwhile, in countries with nationally integrated health systems, citizens may be able to assemble their data across different institutions by using a secure server, such as Azure [31]. Henry et al [18] created a real-time prediction system for critically ill patients that alerts staff to elevated sepsis risk and tracks trends in vitals by using cloud-based technology. In the outpatient setting, Kawamoto et al [30] incorporated data from several EHRs into an existing risk prediction model.

A total of 3 studies described visualization tools. Gordon et al [29] generated visual aids to show patients' thrombocytopenia trends, along with computer-generated calculated panel reactive antibody levels, to facilitate the judicious use of platelet transfusions by physicians and blood banks, and Thayer et al [36] used translated FHIR concepts to graphically display a patient's asthma history within a chart. Xiao et al [41] were able to use knowledge graph ontologies to map FHIR and Observational Medical Outcomes Partnership (OMOP) data standards.

Despite the considerable benefits of cloud-based systems, they can present additional security challenges. These range from traditional cybersecurity problems (including problems related to data security, access control, and the transmission of data over a network) to more CDSS-specific concerns (such as privacy leakage, whereby models can be queried by outside parties). HL7 FHIR has put forward specific security protocols in response to safety concerns, including the use of secure http communication channels, open authorization, and provenance (documentation of the origin, possession, and history of a piece of data) techniques, among others [42].

Data Management and Analytic Platforms

The rise in computing power and distributed system technologies facilitates general-purpose platforms that provide data standardization, data analysis, and model integration. Of the 39 included articles, 10 described FHIR-compliant data management and analytic platforms, as listed in Table 2. Although CDSSs require interoperability and multicenter clinical implementation, many clinical platforms did not support the real-time data integration that is necessary for clinical adoption.

Table . Summary of interoperable data management and analytic platforms.

Source, year	Functionalities	Strengths	Limitations
Gruendner et al [14], 2019	Data analysis and model deployment in clinical environments	Applied Docker virtualization that facilitates deployment across different environments	Poor performance on Extract, Transform, Load processing; relatively inefficient (bottleneck) FHIR[a] transformation; and does not support real-time data processing
Haarbrandt et al [24], 2018	Integrating and transforming health data for oncology, cardiology, and infection control	Open-source platform that allows for patient-level data sharing	Does not support real-time data processing
Helm et al [43], 2022	Builds interoperability between FHIR and BPMN[b]	Supports BPMN clinical process models and improves explainability	Lacks some functionalities of the systems when used independently
Khalilia et al [25], 2015	Clinical predictive modeling using web services via HL7[c] FHIR standards	Maintains good performance across many different algorithms	Does not support real-time data processing
Kopanitsa [44], 2019	Connects multiple health data systems	Has clear, effective workflows	Does not support real-time data processing
Marteau et al [45], 2022	Increases availability of clinical pediatric data using OMOP[d] on FHIR	Implementation across multiple local environments	Not yet tested on real-word applications
Metke-Jimenez et al [46], 2018	Data searching, upgrading, and analyzing within multiple concept and category maps.	Syndication models automatically update the data	Does not support real-time data processing
Semenov et al [47], 2019	Clinical predictive analytics with text outputs to physicians and patients	Produces free-text outputs and graph visualizations pertaining to model recommendations	Limited support for real-time data processing.
Thiess et al [17], 2022	Application for support of shared decision-making in context of drug-drug interactions	Embedded interoperability functions within modular CDSS[e] architectures	Performance testing limited to electronic health record training module
Xiao et al [41], 2022	Enables FHIR and OMOP interoperability with generated clinical knowledge graphs	Semantic foundation for development of explainable tools	Future iterations will require expansion of mapping systems

[a]FHIR: Fast Healthcare Interoperability Resources.

[b]BPMN: Business Process Model and Notation 2.0.

[c]HL7: Health Level 7.

[d]OMOP: Observational Medical Outcomes Partnership.

[e]CDSS: clinical decision support system.

Several papers addressed the challenge of integrating data from heterogeneous sources. Haarbrandt et al [24] proposed a platform that addresses this problem by developing techniques for converting disparate sources to FHIR standards prior to integration. The system is protected via fine-grained use-and-access control mechanisms that ensure secure data transmission among participating data sources. Metke-Jimenez et al [46] proposed an alternative approach to integrating several ontologies into a single web ontology language, allowing for updates to the ontology without changing the underlying data. For example, one could update the definition of *sepsis* and readily find all patients meeting the new definition. Distributed processing systems can be further enhanced via compartmentalization. Kopanitsa [44] and Semenov et al [47] developed a microservice platform that connects multiple systems via FHIR APIs. This platform was used to successfully deploy 400 CDSS models and 128 Bayesian diagnostic models in real time. Important to precision medicine, genomics data can now be linked to FHIR clinical data; 2 groups have created interoperability between the Variant Call Format for next-generation sequencing and FHIR [13,14].

Clinical information systems can aid in medical research, if properly designed. Although a prototype system proposed by Khalilia et al [25] ran 9 different machine learning models to generate data-driven, patient-level predictions, it lacked a researcher interface for the development and training of new models. In contrast, the KETOS platform proposed by Gruendner et al [14] allows researchers to request data sets, define cohorts, develop models, and deploy them as a web service. Both systems use Extract, Transform, Load pipelines to convert EHR data from their native format to the OMOP common data model format before storage. The KETOS platform's comprehensive approach to data management and model deployment can aid researchers with limited backgrounds in data science.

Auxiliary Modules and APIs

Artificial intelligence clinical information systems depend on robust and secure APIs to interact with the clinical environment. APIs define quality and security standards for each type of interaction with external systems (eg, EHR systems, web browsers, and medical devices). Article summaries are shown in Table 3.

Table . Summary of auxiliary modules and application programming interfaces (APIs).

Source, year	Functionalities	Strengths	Limitations
Altamimi [23], 2016	Provide security for FHIR[a] functions to ensure patients' privacy	Policies can be adjusted for circumstances (eg, emergency medical conditions can override privacy constructs)	There is no description of a user-side module, which would be necessary for clinical application
Alterovitz et al [48], 2015	Link clinical and genomic data with an FHIR-compliant API for clinical decision support	Ensures consistent semantics in clinical data and handles multiple types of genomic data	Effects of clinical decision support apps on decision-making and outcomes were not reported.
Dolin et al [49], 2021	Variant Call Format–to–FHIR genomic standard converter	Readily deployable to CDSS[b]	Limited independent data analysis and does not support real-time data processing
Kasparick et al [50], 2019	Model an FHIR-compliant protocol for artificial intelligence–based systems	Supports multiple devices and multiple domains of data	No clinical testing
Kopanitsa and Ivanov [51], 2018	FHIR-compliant APIs for data modeling	High data exchanging efficiency	No clinical testing
Gabetta et al [52], 2021	FHIR-on-OMOP[c] platform to support data storage and retrieval	Use of standard OMOP vocabularies	No clinical testing
Guinez-Molinos et al [53], 2021	Reports COVID-19 test results to central authority	Interoperable and portable; functionally verified with a pilot study	Developed using a predecessor system
Mandel et al [26], 2016	Updating an API platform with FHIR standards	Improves API interoperability	Establishes feasibility, but effects on clinical decision-making and outcomes are unknown.
Rafee et al [54], 2022	LOINC[d]-mapped core data set for eligibility screening	Rapid EHR[e] screening for patient recruitment	Relied on expert labeling, which limits scalability
Wood et al [55], 2021	Allows sharing of patient data among care provision sites for hematologic disorders	Compatible across EHRs	Framework alone; awaiting evidence of implementation
Yoo et al [56], 2022	Method for integrating CDSS applications with EHR	Transformation of EHR data into FHIR format for input into a reasoning engine	No validation of performance indices and usability of tested models

[a]FHIR: Fast Healthcare Interoperability Resources.

[b]CDSS: clinical decision support system.

[c]OMOP: Observational Medical Outcomes Partnership.

[d]LOINC: Logical Observation Identifiers Names and Codes.

[e]EHR: electronic health record.

Of the included articles, 5 described auxiliary modules and APIs. Mandel et al [26] applied FHIR standards to the SMART platform, improving its interoperability by providing standard authentication, authorization, and profiling. The prototype genomics standard developed by Alterovitz et al [48], meanwhile, is currently in trial use to facilitate the consistent integration of clinical and genomic information through SMART-on-FHIR application. The application developers found the FHIR v4.0.1 specification easy to leverage, even without prior experience with FHIR.

Although FHIR has predefined resources and mechanisms for transmitting orders and values, methods for creating and validating orders are not predefined. To address this issue, Kopanitsa and Ivanov [51] proposed an FHIR-based mechanism for integrating laboratory and hospital information systems. The system generated laboratory orders, using the available tests in the laboratory information system, and prompted the user for relevant information (such as how many laboratory samples should be collected and when they should be collected). It is challenging to make clinical information systems both highly interoperable and secure without compromising data workflows. SecFHIR is an XML-based security approach to FHIR resources. Using schema permissions built into XML documents, Altamimi [23] generated robust security profiles that were context-aware (eg, privacy constraints can be overridden in emergency care situations).

Timely data availability is another barrier to implementing CDSSs in high-acuity environments. Kasparick et al [50]

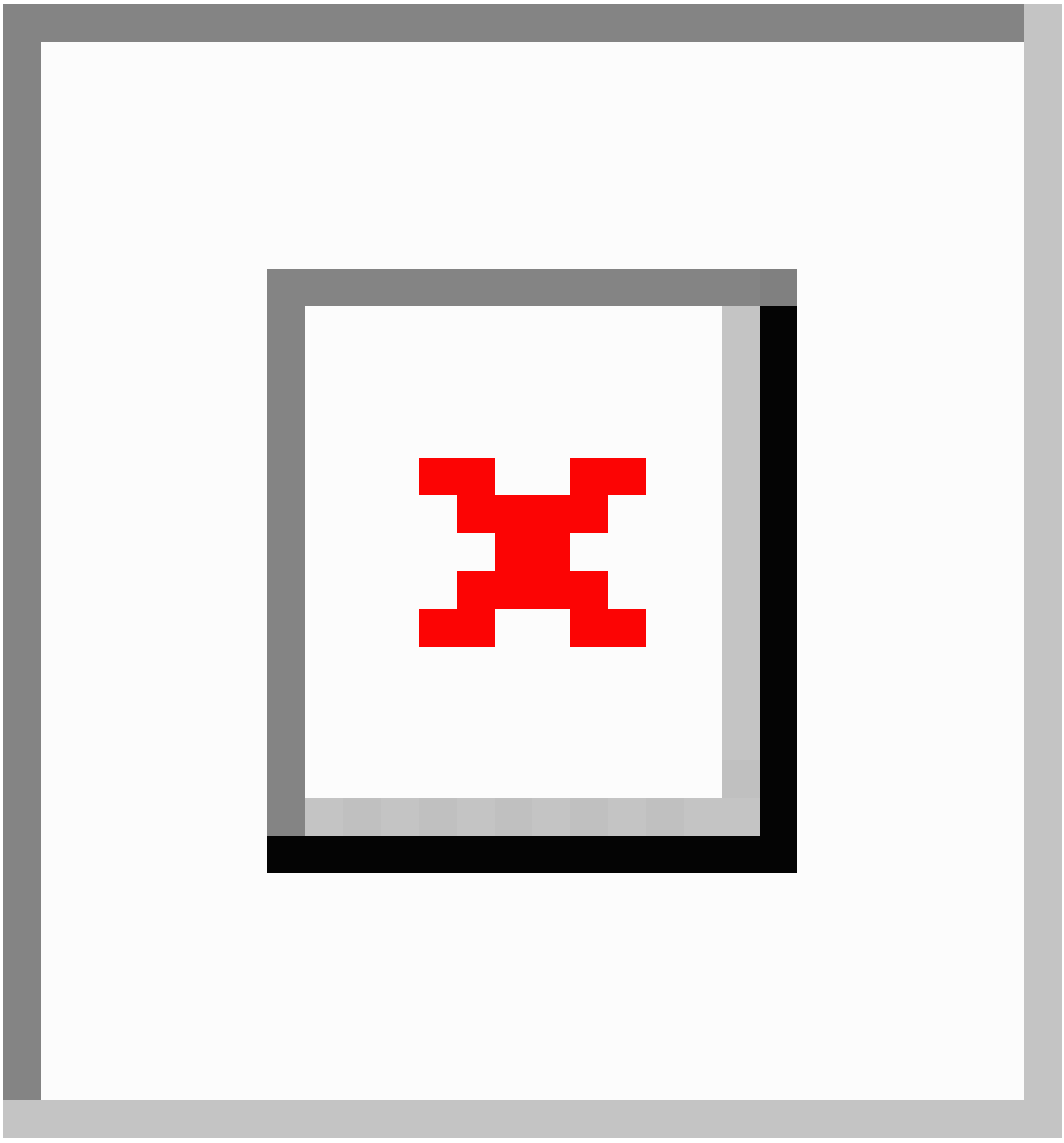
proposed a reference model to address the timeliness challenge by connecting medical devices to FHIR servers. This approach allows the APIs to function as data sources for predictive analytic and decision support systems. By using these methods, clinical information systems can maintain high interoperability and security without compromising data workflow. This has allowed for the development of disease-specific data hubs, which facilitate research on rare conditions or for reporting the results of COVID-19 polymerase chain reaction tests from disparate testing sites to a central authority [53,55]. CDS hooks are another technology that permit the integration of EHR data into external health care applications [57]. Used in collaboration with SMART on FHIR, CDS hooks are triggered by a specific action within the EHR (ie, ordering a medication). The CDS hooks then link the corresponding EHR data to an environment of decision support applications [58]. These CDS applications can then push recommendations in the form of "CDS cards" to the clinician. These technologies are currently being tested in real-word settings [59,60].

Discussion

Key Findings

Although significant progress has been made in the field of FHIR data standards, this scoping review demonstrates that most CDSSs lack interoperability and actionable content. Several modules and APIs demonstrate the potential to enhance these systems, but they were not comprehensively integrated into the existing clinical workflows or were not validated on external patient populations. These limitations collectively reveal several opportunities to improve on existing methods to produce ideal clinical information systems, as illustrated in Figure 1.

Figure 1. Sample model of a proposed machine learning–enabled clinical information system using FHIR data standards. AI: artificial intelligence; API: application programming interface; FHIR: Fast Healthcare Interoperability Resources; HL7: Health Level 7; IoT: Internet of Things; OMOP: Observational Medical Outcomes Partnership.



Foundational Infrastructures Tailored to Individual Needs

Ideally, clinical information systems would function as innovation hubs for patient care and health care research. Due to the proprietary nature of hardware and software systems in institutions, infrastructure components (eg, data transformation, model development, authentication, and monitoring) are often painstakingly created de novo. Platforms, such as KETOS, however, can enable the sharing of core infrastructure, greatly accelerating the development and deployment of applications that are tailored to the needs of individual researchers, groups, and projects [14]. This "health care application development hub" would be shared for different applications to reuse models and for data processing and analyzing services.

Facilitating Interoperability Among Data Systems

Interoperability represents the goal of successful, cross-institutional sharing of data without additional, special effort. This remains in contrast to the current environment of fragmented data systems. Several elements of the current system impede progress toward integration and should be addressed. Sources of patient health information are numerous. At the point of data collection, clinicians may opt to store information in separate departments, erroneously duplicate patient descriptors, preferentially format or describe data, or use older data standards

(HL7 v2 and v3). Further complexity may arise from the use of siloed systems, as exemplified by legacy systems built with local, stand-alone data conventions and incompatible ontologies [61]. The road to full interoperability is therefore paved with standards built to define, represent, transfer, and protect data as they travel between actors. Common communications standards, such as FHIR, have provided a useful framework for standardizing data transmission while maintaining semantic integrity at the patient level [15,56]. Importantly, these standards are built to mobilize data from legacy systems, making closely held data more publicly available [17]. By using OMOP-on-FHIR algorithms, pediatric data from Shriners Hospitals for Children can now be shared more widely by researchers [45]. More recently, 2 studies have examined the use of deep learning and transformer techniques to convert data elements in the EHR to interoperable FHIR standards, with subsequent application in prediction models [21,62]. Automation in data capture has the potential to reduce the costs and time associated with manual extraction.

Overcoming Organizational Resistance to Interoperability Standards

Despite the benefits of an interoperable health data ecosystem, stakeholders are rarely incentivized to implement data standards. Organizational resistance to interoperability may stem from cultural differences, unfamiliarity with new technologies, or the fear that a newly adopted information-sharing standard may quickly become obsolete [63,64]. Among organizations, concerns regarding the loss of autonomy, a lack of trust, and the failure to realize financial gains impede interoperability and lead to so-called "information blocking." The policies contained within the 21st Century Cures Act aim to improve information flow among actors in the system [65,66]. Apple, Google, and Samsung now have patient-facing health records that were developed along with FHIR standards to comply with these policies. In addition, while implementation models exist to help streamline the adoption of CDSSs, they contain important methodical flaws [67].

Hiring Specialists to Manage Standards Adoption

Unfamiliarity with interoperability standards may represent a substantial hurdle to adoption and subsequent interoperability. This challenge creates demand for subject matter experts who are familiar with the architecture, function, and implementation of data standards. Such experts must be able to anticipate the specific challenges of adapting their particular legacy systems to the interoperable standard but also recognize the benefit of successful adoption to guide organizational buy-in [68].

Timely Data Acquisition

The need for timeliness in data sharing is driven both by data availability and by opportunities for real-time treatment support. An obvious example of this can be seen with continuous glucose monitoring units for patients with diabetes, which provide a regular source of data that can be implemented immediately to adjust insulin therapy [69,70].

System scalability is also essential to this task. Many of the systems evaluated in this review cannot scale in real time, as data volume or velocity increases dynamically (eg, processing 1000 patients in real time vs processing 100 patients in a static, retrospective training cohort). When scalability is impaired, predictions may not be delivered in time to augment clinical decision-making. Health Insurance Portability and Accountability Act (HIPAA)–compliant cloud platforms can scale allocated resources on demand. Therefore, optimal clinical information systems must offer scalability that is commensurate with the expected volume and velocity of data.

Minimizing Discoverable Patient Data

Each institution has policies that comply with municipal and federal security and privacy laws, making it challenging to share and aggregate data across multiple institutions. These challenges have been met with creative methods for aggregating multicenter data while maintaining patient privacy. One such method is to request only the minimum necessary information. This approach is emphasized heavily in the HIPAA and exemplified by El-Sappagh et al [28], who described a system that requests only the required EHR data elements for a specific patient. Other such mechanisms include authorization programs (enables specialized control over access to patient data), https, and WebSockets (Internet Engineering Task Force; provides secure communication over networks).

Alternatively, models can benefit from the knowledge derived from other data sets—usually in the form of model gradients or coefficients—without sharing the underlying data. This is known as *federated learning*—a system that trains on many local models with the same architecture and then aggregates the knowledge derived from each center into a global model (Figure 1). Although such an approach greatly reduces security and privacy risks by keeping the source records under the control of each local institution, even the gradients themselves pose a minor risk due to privacy leakage [28,71-75]. This risk, however, can be further reduced via the automated obfuscation of high-risk records or by adding noise to the gradients and coefficients before transmitting them to the central model. Given these advantages, federated learning is poised to supplant other methods for ensuring the data security and privacy of clinical information systems.

Finally, the recent explosion of large language models has raised further concerns regarding data privacy, as they are trained on clinical notes. This is an active field of study with multiple avenues for further research [76,77].

Conclusions

Machine learning–enabled clinical analytic and decision support systems have the potential to improve health care by automating standardized workflows and augmenting clinical decision-making. Nevertheless, most CDSSs lack interoperability and evidence of clinical utility. Common data models and interoperable data management platforms can address these limitations, but most intelligent clinical platforms are also compromised by the inadequate scalability for supporting real-time data processing. Existing clinical information systems could be improved by using foundational code infrastructures, common data models, and secure data processing and analytics on real-time platforms. Further progress in implementing these elements can generate information

systems that improve care by helping patients, caregivers, and clinicians make effective, well-informed clinical decisions.

Acknowledgments

We would like to thank Zachary Hodges, Zhang Feng, and Shounak Datta for their initial contributions.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist for scoping reviews.
[PDF File, 599 KB - medinform_v11i1e48297_app1.pdf]

Multimedia Appendix 2
Search criteria.
[DOCX File, 12 KB - medinform_v11i1e48297_app2.docx]

References

1. Centers for Disease Control and Prevention, National Center for Health Statistics. Early release of selected estimates based on data from the 2018 National Health Interview Survey. URL: www.cdc.gov/nchs/nhis/releases/released201905.htm [accessed 2023-07-28]
2. Birkhead GS, Klompas M, Shah NR. Uses of electronic health records for public health surveillance to advance public health. Annu Rev Public Health 2015 Mar 18;36:345-359. [doi: 10.1146/annurev-publhealth-031914-122747] [Medline: 25581157]
3. Adler-Milstein J, Holmgren AJ, Kralovec P, Worzala C, Searcy T, Patel V. Electronic health record adoption in US hospitals: the emergence of a digital "advanced use" divide. J Am Med Inform Assoc 2017 Nov 1;24(6):1142-1148. [doi: 10.1093/jamia/ocx080] [Medline: 29016973]
4. Stanford Medicine. Stanford Medicine 2017 health trends report: harnessing the power of data in health. 2017. URL: med.stanford.edu/content/dam/sm/sm-news/documents/StanfordMedicineHealthTrendsWhitePaper2017.pdf [accessed 2023-07-28]
5. Unni N, Peddinghaus M, Tormey CA, Stack G. Record fragmentation due to transfusion at multiple health care facilities: a risk factor for delayed hemolytic transfusion reactions. Transfusion 2014 Jan;54(1):98-103. [doi: 10.1111/trf.12251] [Medline: 23711236]
6. Hempstead K, Delia D, Cantor JC, Nguyen T, Brenner J. The fragmentation of hospital use among a cohort of high utilizers: implications for emerging care coordination strategies for patients with multiple chronic conditions. Med Care 2014 Mar;52(Suppl 3):S67-S74. [doi: 10.1097/MLR.0000000000000049] [Medline: 24561761]
7. Justiniano CF, Xu Z, Becerra AZ, Aquina CT, Boodry CI, Swanger AA, et al. Surgeon care fragmentation during readmission after colorectal surgery is associated with increased mortality: continuity of care counts. J Am Coll Surg 2017;225(4):S126-S127. [doi: 10.1016/j.jamcollsurg.2017.07.280]
8. Tsai TC, Orav EJ, Jha AK. Care fragmentation in the postdischarge period: surgical readmissions, distance of travel, and postoperative mortality. JAMA Surg 2015 Jan;150(1):59-64. [doi: 10.1001/jamasurg.2014.2071] [Medline: 25472595]
9. HL7 International. FHIR V5.0.0. URL: hl7.org/fhir [accessed 2023-07-28]
10. Shortliffe EH, Cimino JJ. Biomedical Informatics: Computer Applications in Health Care and Biomedicine. Cham, Switzerland: Springer; 2021. [doi: 10.1007/978-3-030-58721-5]
11. Payne TH, Corley S, Cullen TA, Gandhi TK, Harrington L, Kuperman GJ, et al. Report of the AMIA EHR-2020 Task Force on the status and future direction of EHRs. J Am Med Inform Assoc 2015 Sep;22(5):1102-1110. [doi: 10.1093/jamia/ocv066] [Medline: 26024883]
12. Griffin AC, He L, Sunjaya AP, King AJ, Khan Z, Nwadiugwu M, et al. Clinical, technical, and implementation characteristics of real-world health applications using FHIR. JAMIA Open 2022 Oct 12;5(4):ooac077. [doi: 10.1093/jamiaopen/ooac077] [Medline: 36247086]
13. Dolin RH, Boxwala A, Shalaby J. A pharmacogenomics clinical decision support service based on FHIR and CDS hooks. Methods Inf Med 2018 Dec;57(S 02):e115-e123. [doi: 10.1055/s-0038-1676466] [Medline: 30605914]
14. Gruendner J, Schwachhofer T, Sippl P, Wolf N, Erpenbeck M, Gulden C, et al. KETOS: clinical decision support and machine learning as a service - A training and deployment platform based on Docker, OMOP-CDM, and FHIR web services. PLoS One 2019 Oct 3;14(10):e0223010. [doi: 10.1371/journal.pone.0223010] [Medline: 31581246]

15. Curran RL, Kukhareva PV, Taft T, Weir CR, Reese TJ, Nanjo C, et al. Integrated displays to improve chronic disease management in ambulatory care: a SMART on FHIR application informed by mixed-methods user testing. J Am Med Inform Assoc 2020 Aug 1;27(8):1225-1234. [doi: 10.1093/jamia/ocaa099] [Medline: 32719880]
16. Gaebel J, Cypko MA, Lemke HU. Accessing patient information for probabilistic patient models using existing standards. Stud Health Technol Inform 2016;223:107-112. [Medline: 27139392]]
17. Thiess H, Del Fiol G, Malone DC, Cornia R, Sibilla M, Rhodes B, et al. Coordinated use of Health Level 7 standards to support clinical decision support: case study with shared decision making and drug-drug interactions. Int J Med Inform 2022 Mar 21;162:104749. [doi: 10.1016/j.ijmedinf.2022.104749] [Medline: 35358893]
18. Henry JR, Lynch D, Mals J, Shashikumar SP, Holder A, Sharma A, et al. A FHIR-enabled streaming sepsis prediction system for ICUs. Annu Int Conf IEEE Eng Med Biol Soc 2018 Jul;2018:4093-4096. [doi: 10.1109/EMBC.2018.8513347] [Medline: 30441256]
19. Wang S, Han J, Jung SY, Oh TJ, Yao S, Lim S, et al. Development and implementation of patient-level prediction models of end-stage renal disease for type 2 diabetes patients using fast healthcare interoperability resources. Sci Rep 2022 Jul 4;12(1):11232. [doi: 10.1038/s41598-022-15036-6] [Medline: 35789173]
20. Semenov I, Kopanitsa G. Decision support system based on FHIR profiles. Stud Health Technol Inform 2018;249:117-121. [Medline: 29866966]]
21. Sun H, Depraetere K, Meesseman L, De Roo J, Vanbiervliet M, De Baerdemaeker J, et al. A scalable approach for developing clinical risk prediction applications in different hospitals. J Biomed Inform 2021 Jun;118:103783. [doi: 10.1016/j.jbi.2021.103783] [Medline: 33887456]
22. Iglesias N, Juarez JM, Campos M. Comprehensive analysis of rule formalisms to represent clinical guidelines: selection criteria and case study on antibiotic clinical guidelines. Artif Intell Med 2020 Mar;103:101741. [doi: 10.1016/j.artmed.2019.101741] [Medline: 31928849]
23. Altamimi AM. SecFHIR: a security specification model for fast healthcare interoperability resources. Int J Adv Comput Sci Appl 2016 Jun;7(6):350-355. [doi: 10.14569/IJACSA.2016.070645]
24. Haarbrandt B, Schreiweis B, Rey S, Sax U, Scheithauer S, Rienhoff O, et al. HiGHmed - an open platform approach to enhance care and research across institutional boundaries. Methods Inf Med 2018 Jul;57(S 01):e66-e81. [doi: 10.3414/ME18-02-0002] [Medline: 30016813]
25. Khalilia M, Choi M, Henderson A, Iyengar S, Braunstein M, Sun J. Clinical predictive modeling development and deployment through FHIR web services. AMIA Annu Symp Proc 2015 Nov 5;2015:717-726. [Medline: 26958207]]
26. Mandel JC, Kreda DA, Mandl KD, Kohane IS, Ramoni RB. SMART on FHIR: a standards-based, interoperable apps platform for electronic health records. J Am Med Inform Assoc 2016 Sep;23(5):899-908. [doi: 10.1093/jamia/ocv189] [Medline: 26911829]
27. Hong N, Wen A, Shen F, Sohn S, Wang C, Liu H, et al. Developing a scalable FHIR-based clinical data normalization pipeline for standardizing and integrating unstructured and structured electronic health record data. JAMIA Open 2019 Oct 18;2(4):570-579. [doi: 10.1093/jamiaopen/ooz056] [Medline: 32025655]
28. El-Sappagh S, Ali F, Hendawi A, Jang JH, Kwak KS. A mobile health monitoring-and-treatment system based on integration of the SSN sensor ontology and the HL7 FHIR standard. BMC Med Inform Decis Mak 2019 May 10;19(1):97. [doi: 10.1186/s12911-019-0806-z] [Medline: 31077222]
29. Gordon WJ, Baronas J, Lane WJ. A FHIR human leukocyte antigen (HLA) interface for platelet transfusion support. Appl Clin Inform 2017 Jun 7;8(2):603-611. [doi: 10.4338/ACI-2017-01-CR-0010] [Medline: 28850154]
30. Kawamoto K, Kukhareva PV, Weir C, Flynn MC, Nanjo CJ, Martin DK, et al. Establishing a multidisciplinary initiative for interoperable electronic health record innovations at an academic medical center. JAMIA Open 2021 Jul 31;4(3):ooab041. [doi: 10.1093/jamiaopen/ooab041] [Medline: 34345802]
31. Park C, You SC, Jeon H, Jeong CW, Choi JW, Park RW. Development and validation of the Radiology Common Data Model (R-CDM) for the international standardization of medical imaging data. Yonsei Med J 2022 Jan;63(Suppl):S74-S83. [doi: 10.3349/ymj.2022.63.S74] [Medline: 35040608]
32. Schleyer T, Williams L, Gottlieb J, Weaver C, Saysana M, Azar J, et al. The Indiana Learning Health System Initiative: early experience developing a collaborative, regional learning health system. Learn Health Syst 2021 Jun 23;5(3):e10281. [doi: 10.1002/lrh2.10281] [Medline: 34277946]
33. Semenov I, Kopanitsa G, Denisov D, Alexandr Y, Osenev R, Andreychuk Y. Patients decision aid system based on FHIR profiles. J Med Syst 2018 Jul 31;42(9):166. [doi: 10.1007/s10916-018-1016-4] [Medline: 30066031]
34. Séroussi B, Guézennec G, Lamy JB, Muro N, Larburu N, Sekar BD, et al. Reconciliation of multiple guidelines for decision support: a case study on the multidisciplinary management of breast cancer within the DESIREE project. AMIA Annu Symp Proc 2018 Apr 16;2017:1527-1536. [Medline: 29854222]]
35. Tarumi S, Takeuchi W, Chalkidis G, Rodriguez-Loya S, Kuwata J, Flynn M, et al. Leveraging artificial intelligence to improve chronic disease care: methods and application to pharmacotherapy decision support for type-2 diabetes mellitus. Methods Inf Med 2021 Jun;60(S 01):e32-e43. [doi: 10.1055/s-0041-1728757] [Medline: 33975376]
36. Thayer JG, Ferro DF, Miller JM, Karavite D, Grundmeier RW, Utidjian L, et al. Human-centered development of an electronic health record-embedded, interactive information visualization in the emergency department using fast healthcare

interoperability resources. J Am Med Inform Assoc 2021 Jul 14;28(7):1401-1410. [doi: 10.1093/jamia/ocab016] [Medline: 33682004]
37. Wang Z, Song M, Zhang Z, Song Y, Wang Q, Qi H. Beyond Inferring class representatives: user-level privacy leakage from Federated learning. Presented at: IEEE INFOCOM 2019 - IEEE Conference on Computer Communications; Paris, France p. 2512-2520. [doi: 10.1109/INFOCOM.2019.8737416]
38. Whitaker B, Pizarro J, Deady M, Williams A, Ezzeldin H, Belov A, et al. Detection of allergic transfusion-related adverse events from electronic medical records. Transfusion 2022 Oct;62(10):2029-2038. [doi: 10.1111/trf.17069] [Medline: 36004803]
39. HL7 International. Minimal Common Oncology Data Elements (mCODE) implementation guide. URL: hl7.org/fhir/us/mcode [accessed 2023-07-31]
40. HL7 International. CodeX. URL: www.hl7.org/codex [accessed 2023-07-31]
41. Xiao G, Pfaff E, Prud'hommeaux E, Booth D, Sharma DK, Huo N, et al. FHIR-Ontop-OMOP: building clinical knowledge graphs in FHIR RDF with the OMOP Common Data Model. J Biomed Inform 2022 Oct;134:104201. [doi: 10.1016/j.jbi.2022.104201] [Medline: 36089199]
42. HL7 International. 6.1.0 FHIR security. URL: build.fhir.org/security.html [accessed 2023-07-31]
43. Helm E, Pointner A, Krauss O, Schuler A, Traxler B, Arthofer K, et al. FHIR2BPMN: delivering actionable knowledge by transforming between clinical pathways and executable models. Stud Health Technol Inform 2022 May 16;292:9-14. [doi: 10.3233/SHTI220311] [Medline: 35575842]
44. Kopanitsa G. Microservice architecture to provide medical data management for decision support. Stud Health Technol Inform 2019;261:230-235. [Medline: 31156121]]
45. Marteau BL, Zhu Y, Giuste F, Shi W, Carpenter A, Hilton C, et al. Accelerating multi-site health informatics with streamlined data infrastructure using OMOP-on-FHIR. July 11-15, 2022 Presented at: 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC); Glasgow, Scotland.
46. Metke-Jimenez A, Steel J, Hansen D, Lawley M. Ontoserver: a syndicated terminology server. J Biomed Semantics 2018 Sep 17;9(1):24. [doi: 10.1186/s13326-018-0191-z] [Medline: 30223897]
47. Semenov I, Osenev R, Gerasimov S, Kopanitsa G, Denisov D, Andreychuk Y. Experience in developing an FHIR medical data management platform to provide clinical decision support. Int J Environ Res Public Health 2019 Dec 20;17(1):73. [doi: 10.3390/ijerph17010073] [Medline: 31861851]
48. Alterovitz G, Warner J, Zhang P, Chen Y, Ullman-Cullere M, Kreda D, et al. SMART on FHIR genomics: facilitating standardized clinico-genomic apps. J Am Med Inform Assoc 2015 Nov;22(6):1173-1178. [doi: 10.1093/jamia/ocv045] [Medline: 26198304]
49. Dolin RH, Gothi SR, Boxwala A, Heale BSE, Husami A, Jones J, et al. vcf2fhir: a utility to convert VCF files into Hl7 FHIR format for genomics-EHR integration. BMC Bioinformatics 2021 Mar 2;22(1):104. [doi: 10.1186/s12859-021-04039-1] [Medline: 33653260]
50. Kasparick M, Andersen B, Franke S, Rockstroh M, Golatowski F, Timmermann D, et al. Enabling artificial intelligence in high acuity medical environments. Minim Invasive Ther Allied Technol 2019 Apr;28(2):120-126. [doi: 10.1080/13645706.2019.1599957] [Medline: 30950665]
51. Kopanitsa G, Ivanov A. Implementation of Fast Healthcare Interoperability Resources for an integration of laboratory and hospital information systems. Stud Health Technol 2018;247:11-15. [Medline: 29677913]]
52. Gabetta M, Alloni A, Polce F, Lanzola G, Parimbelli E, Barbarini N. Development of a FHIR layer on top of the OMOP Common Data Model for the CAPABLE project. Stud Health Technol Inform 2021 Nov 18;287:28-29. [doi: 10.3233/SHTI210804] [Medline: 34795073]
53. Guinez-Molinos S, Andrade JM, Negrete AM, Vidal SE, Rios E. Interoperable platform to report polymerase chain reaction SARS-CoV-2 tests from laboratories to the Chilean government: development and implementation study. JMIR Med Inform 2021 Jan 20;9(1):e25149. [doi: 10.2196/25149] [Medline: 33417587]
54. Rafee A, Riepenhausen S, Neuhaus P, Meidt A, Dugas M, Varghese J. ELaPro, a LOINC-mapped core dataset for top laboratory procedures of eligibility screening for clinical trials. BMC Med Res Methodol 2022 May 14;22(1):141. [doi: 10.1186/s12874-022-01611-y] [Medline: 35568796]
55. Wood WA, Marks P, Plovnick RM, Hewitt K, Neuberg DS, Walters S, et al. ASH Research Collaborative: a real-world data infrastructure to support real-world evidence development and learning healthcare systems in hematology. Blood Adv 2021 Dec 14;5(23):5429-5438. [doi: 10.1182/bloodadvances.2021005902] [Medline: 34673922]
56. Yoo J, Lee J, Min JY, Choi SW, Kwon JM, Cho I, et al. Development of an interoperable and easily transferable clinical decision support system deployment platform: system design and development study. J Med Internet Res 2022 Jul 27;24(7):e37928. [doi: 10.2196/37928] [Medline: 35896020]
57. CDS hooks. URL: cds-hooks.org [accessed 2023-07-31]
58. Strasberg HR, Rhodes B, Del Fiol G, Jenders RA, Haug PJ, Kawamoto K. Contemporary clinical decision support standards using Health Level Seven International Fast Healthcare Interoperability Resources. J Am Med Inform Assoc 2021 Jul 30;28(8):1796-1806. [doi: 10.1093/jamia/ocab070] [Medline: 34100949]

59. Jung S, Bae S, Seong D, Oh OH, Kim Y, Yi BK. Shared Interoperable clinical decision support service for drug-allergy interaction checks: implementation study. JMIR Med Inform 2022 Nov 10;10(11):e40338. [doi: 10.2196/40338] [Medline: 36355401]
60. Morgan KL, Kukhareva PV, Warner PB, Wilkof J, Snyder M, Horton D, et al. Using CDS hooks to increase SMART on FHIR app utilization: a cluster-randomized trial. J Am Med Inform Assoc 2022 Aug 16;29(9):1461-1470. [doi: 10.1093/jamia/ocac085] [Medline: 35641136]
61. Olaronke I, Soriyan A, Gambo I, Olaleke J. Interoperability in healthcare: benefits, challenges and resolutions. Int J Innov Appl Stud 2013 May;3(1):262-270.
62. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med 2018 May 8;1:18. [doi: 10.1038/s41746-018-0029-1] [Medline: 31304302]
63. Han L, Liu J, Evans R, Song Y, Ma J. Factors influencing the adoption of health information standards in health care organizations: a systematic review based on best fit framework synthesis. JMIR Med Inform 2020 May 15;8(5):e17334. [doi: 10.2196/17334] [Medline: 32347800]
64. Henning F. Adoption of Interoperability standards in government information networks: an initial framework of influence factors. Presented at: ICEGOV '13: 7th International Conference on Theory and Practice of Electronic Governance; Seoul Republic of Korea p. 264-267. [doi: 10.1145/2591888.2591936]
65. Everson J, Patel V, Adler-Milstein J. Information blocking remains prevalent at the start of 21st Century Cures Act: results from a survey of health information exchange organizations. J Am Med Inform Assoc 2021 Mar 18;28(4):727-732. [doi: 10.1093/jamia/ocaa323] [Medline: 33410891]
66. Health and Human Services Department. 21st Century Cures Act: Interoperability, information blocking, and the ONC Health IT Certification Program. 2020. URL: www.federalregister.gov/documents/2020/05/01/2020-07419/21st-century-cures-act-interoperability-information-blocking-and-the-onc-health-it-certification [accessed 2023-07-31]
67. Elwyn G, Scholl I, Tietbohl C, Mann M, Edwards AGK, Clay C, et al. "Many miles to go …": a systematic review of the implementation of patient decision support interventions into routine clinical practice. BMC Med Inform Decis Mak 2013 Nov 29;13 Suppl 2(Suppl 2):S14. [doi: 10.1186/1472-6947-13-S2-S14] [Medline: 24625083]
68. Lamprinakos GC, Mousas AS, Kapsalis AP, Kaklamani DI, Venieris IS, Boufis AD, et al. Using FHIR to develop a healthcare mobile application. November 3-5, 2014 Presented at: 2014 4th International Conference on Wireless Mobile Communication and Healthcare - Transforming Healthcare Through Innovations in Mobile and Wireless Technologies (MOBIHEALTH); Athens, Greece. [doi: 10.1109/MOBIHEALTH.2014.7015927]
69. Eberle C, Stichling S, Löhnert M. Diabetology 4.0: scoping review of novel insights and possibilities offered by digitalization. J Med Internet Res 2021 Mar 24;23(3):e23475. [doi: 10.2196/23475] [Medline: 33759789]
70. Hommel E, Olsen B, Battelino T, Conget I, Schütz-Fuhrmann I, Hoogma R, et al. Impact of continuous glucose monitoring on quality of life, treatment satisfaction, and use of medical care resources: analyses from the SWITCH study. Acta Diabetol 2014 Oct;51(5):845-851. [doi: 10.1007/s00592-014-0598-7] [Medline: 25037251]
71. Hitaj B, Ateniese G, Perez-Cruz F. Deep models under the GAN: information leakage from collaborative deep learning. Presented at: CCS '17: 2017 ACM SIGSAC Conference on Computer and Communications Security; Dallas, TX p. 603-618. [doi: 10.1145/3133956.3134012]
72. Kairouz P, McMahan HB, Avent B, Bellet A, Bennis M, Bhagoji AN, et al. Advances and Open Problems in Federated Learning. Norwell, MA: Now Foundations and Trends; 2021. [doi: 10.1561/9781680837896]
73. Nasr M, Shokri R, Houmansadr A. Comprehensive privacy analysis of deep learning: passive and active white-box inference attacks against centralized and federated learning. May 19-23, 2019 Presented at: 2019 IEEE Symposium on Security and Privacy (SP); San Francisco, CA.. [doi: 10.1109/SP.2019.00065]
74. Phong LT, Aono Y, Hayashi T, Wang L, Moriai S. Privacy-preserving deep learning: revisited and enhanced. June 6-7, 2017 Presented at: 8th International Conference on Applications and Techniques in Information Security; Auckland, New Zealand. [doi: 10.1007/978-981-10-5421-1]
75. Wei W, Liu L, Loper M, Chow KH, Gursoy ME, Truex S, et al. A framework for evaluating gradient leakage attacks in federated learning. arXiv. Preprint posted online on April 23, 2020. . [doi: 10.48550/arXiv.2004.10397]
76. Sebastian G. Do ChatGPT and other AI chatbots pose a cybersecurity risk?: an exploratory study. International Journal of Security and Privacy in Pervasive Computing 2023 Jan;15(1):1-11. [doi: 10.4018/IJSPPC.320225]
77. Yoon HJ, Stanley C, Christian JB, Klasky HB, Blanchard AE, Durbin EB, et al. Optimal vocabulary selection approaches for privacy-preserving deep NLP model training for information extraction and cancer epidemiology. Cancer Biomark 2022;33(2):185-198. [doi: 10.3233/CBM-210306] [Medline: 35213361]

Abbreviations

API: application programming interface
CDS: clinical decision support
CDSS: clinical decision support system
EHR: electronic health record

FHIR: Fast Healthcare Interoperability Resources
HIPAA: Health Insurance Portability and Accountability Act
HL7: Health Level 7
LOINC: Logical Observation Identifiers Names and Codes
mCODE: minimal Common Oncology Data Elements
ML-CIS: machine learning–enabled clinical information system
NCIT: National Cancer Institute Thesaurus
OMOP: Observational Medical Outcomes Partnership
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
SMART: Substitutable Medical Applications and Reusable Technologies
SNOMED CT: Systematized Nomenclature of Medicine-Clinical Terms

Edited by C Lovis; submitted 19.04.23; peer-reviewed by G Sakellaropoulos, G Sebastian; revised version received 15.06.23; accepted 17.06.23; published 24.08.23.

Please cite as:
Balch JA, Ruppert MM, Loftus TJ, Guan Z, Ren Y, Upchurch GR, Ozrazgat-Baslanti T, Rashidi P, Bihorac A
Machine Learning–Enabled Clinical Information Systems Using Fast Healthcare Interoperability Resources Data Standards: Scoping Review
JMIR Med Inform 2023;11:e48297
URL: https://medinform.jmir.org/2023/1/e48297
doi:10.2196/48297

Review

Ontologies Applied in Clinical Decision Support System Rules: Systematic Review

Xia Jing[1*], MD, PhD; Hua Min[2*], MD, PhD; Yang Gong[3*], MD, PhD; Paul Biondich[4], MD; David Robinson[5], MD; Timothy Law[6], DO; Christian Nohr[7], PhD; Arild Faxvaag[8], MD, PhD; Lior Rennert[1], PhD; Nina Hubig[9], PhD; Ronald Gimbel[1], PhD

[1]Department of Public Health Sciences, Clemson University, Clemson, SC, United States
[2]College of Public Health, George Mason University, Fairfax, VA, United States
[3]School of Biomedical Informatics, The University of Texas Health Sciences Center at Houston, Houston, TX, United States
[4]Clem McDonald Biomedical Informatics Center, Regenstrief Institute, Indianapolis, IN, United States
[5]Loweswater Consulting, Combria, United Kingdom
[6]Ohio Musculoskeletal and Neurologic Institute, Ohio University, Athens, OH, United States
[7]Department of Planning, Aalborg University, Aalborg, Denmark
[8]Department of Neuromedicine and Movement Science, Norwegian University of Science and Technology, Trondheim, Norway
[9]School of Computing, Clemson University, Clemson, SC, United States
[*]these authors contributed equally

Corresponding Author:
Xia Jing, MD, PhD
Department of Public Health Sciences
Clemson University
519 Edwards Hall
Clemson, SC, 29634
United States
Phone: 1 8646563347
Email: xjing@clemson.edu

Abstract

Background: Clinical decision support systems (CDSSs) are important for the quality and safety of health care delivery. Although CDSS rules guide CDSS behavior, they are not routinely shared and reused.

Objective: Ontologies have the potential to promote the reuse of CDSS rules. Therefore, we systematically screened the literature to elaborate on the current status of ontologies applied in CDSS rules, such as rule management, which uses captured CDSS rule usage data and user feedback data to tailor CDSS services to be more accurate, and maintenance, which updates CDSS rules. Through this systematic literature review, we aim to identify the frontiers of ontologies used in CDSS rules.

Methods: The literature search was focused on the intersection of ontologies; clinical decision support; and rules in PubMed, the Association for Computing Machinery (ACM) Digital Library, and the Nursing & Allied Health Database. Grounded theory and PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 guidelines were followed. One author initiated the screening and literature review, while 2 authors validated the processes and results independently. The inclusion and exclusion criteria were developed and refined iteratively.

Results: CDSSs were primarily used to manage chronic conditions, alerts for medication prescriptions, reminders for immunizations and preventive services, diagnoses, and treatment recommendations among 81 included publications. The CDSS rules were presented in Semantic Web Rule Language, Jess, or Jena formats. Despite the fact that ontologies have been used to provide medical knowledge, CDSS rules, and terminologies, they have not been used in CDSS rule management or to facilitate the reuse of CDSS rules.

Conclusions: Ontologies have been used to organize and represent medical knowledge, controlled vocabularies, and the content of CDSS rules. So far, there has been little reuse of CDSS rules. More work is needed to improve the reusability and interoperability of CDSS rules. This review identified and described the ontologies that, despite their limitations, enable Semantic Web technologies and their applications in CDSS rules.

(JMIR Med Inform 2023;11:e43053) doi:10.2196/43053

KEYWORDS

clinical decision support system rules; clinical decision support systems; interoperability; ontology; Semantic Web technology

Introduction

For more than half a century, clinical decision support systems (CDSSs) have been developed and used in clinical care delivery [1-5]. Some early CDSS examples include Dialog [6], INTERNIST-1 [7-9], Quick Medical Reference [8], and Iliad [10-12]. The effectiveness of CDSSs in clinical care has been established [13-15], with some pioneering researchers' work on CDSS effectiveness particularly noteworthy [16]. Researchers have examined CDSS users' and developers' experiences, discussed their CDSS vision for the future [17], and recommended best practice guidelines in CDSSs [18-22]. Meanwhile, the challenges of CDSSs have been well documented [23]. Meeting clinician information needs is one way a CDSS can help health care providers improve clinical care quality. Many studies, such as Infobutton [13,24], have demonstrated the effectiveness of CDSSs in this aspect. CDSSs are currently routinely used in clinical care, with rates ranging from 68.5% to 100% in primary care settings based in offices [25] in the United States as part of electronic health record (EHR) systems. CDSSs can take many forms, including but not limited to reminders for preventive services (eg, immunizations and screening tests) [26-28], alerts for drug-drug interactions [22,29,30], diagnostic or treatment plan recommendations [31-33], clinician content assistance [34-38], and recommendations for adhering to current clinical practice guidelines [39-41]. CDSSs have played an important role and are widely used in practice to provide safer and better clinical care services.

CDSS rules, which function similarly to the human central nervous system, direct the behaviors of a CDSS during operations by incorporating patient data, contextual information, and medical domain knowledge. The central role of CDSS rules is a decisive factor in the relevance and usefulness of a CDSS in the overall clinical workflow, which impacts whether a CDSS is adopted and routinely used. CDSS rules can be written in Arden syntax [42], Semantic Web Rule Language (SWRL), Jess, Jena, and other programming languages, and the processes are labor intensive. Only specially trained personnel are qualified to write such rules. Moreover, regular updating of CDSS rules is required to keep CDSSs relevant and useful in clinical care delivery. However, the process of developing, updating, and maintaining CDSS rules is time-consuming and resource intensive [4,43], making it difficult for both large institutions and resource-constrained small-scale practices. CDSS rule usage data, such as rule fire rates, overwrite rates, successful rates, and user feedback data, can be collected to improve and customize CDSSs and manage CDSS rules. Typically, CDSS rule maintenance entails adding, deleting, and updating CDSS rules.

Ontologies have been successfully applied to generate and supply domain knowledge in the use, reuse, sharing, and interoperability of information. Ontologies are seen as promising solutions to the challenges of managing and maintaining CDSS rules across institutional boundaries. The Semantic Web is a technology enabled by ontology [44] that is critical in information sharing and reuse [45,46], medicine [47], and CDSSs [48,49]. Although there are numerous definitions of ontology, we used Gruber's definition in this manuscript: "an ontology is a specification of conceptualization" [45]. Interoperability has been identified as a major challenge for health care information technologies, particularly when it comes to sharing health information across institutional or national boundaries. Ontologies have the potential to shorten the interoperability gap.

Reusing and sharing CDSS rules are important, but they are not yet routine operations; thus, we conducted this systematic literature review. This study aims to expand on the current state of using ontologies in CDSS rules by conducting a systematic review of the literature on the intersection of CDSS rules, Semantic Web technologies (particularly ontologies), and use of ontologies in CDSSs. The review is expected to provide a comprehensive view of using ontologies in CDSS rules, with granular details. The results could serve as a basis to form a knowledge framework of the topic that may inspire future research. The research question we intend to answer with this systematic literature review is as follows: What is the current state of using semantic technologies, particularly ontologies, to leverage CDSS rule interoperability? Furthermore, the manually annotated results of selected publications could serve as gold standards for automatically identifying relevant entities in the literature.

Methods

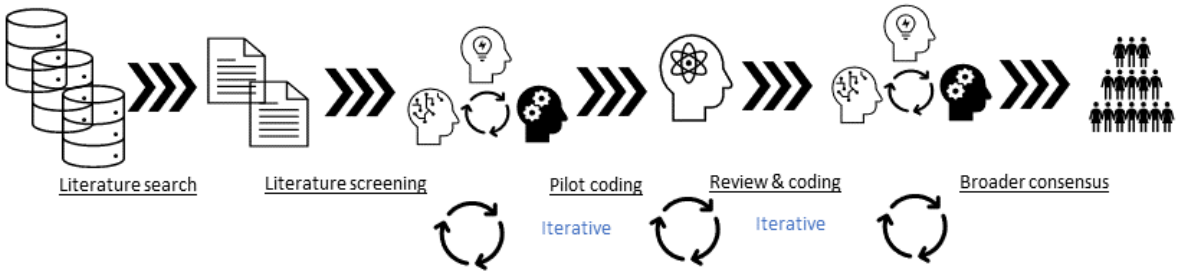
Databases and Search Strategies

Figure 1 illustrates the general workflow we used to conduct this literature review. An initial set of literature searches was conducted on June 2, 2020, which was followed by a review and discussions. The reviewers (XJ, HM, and YG) refined and agreed with the search strategies and searched PubMed, the Association for Computing Machinery (ACM) Digital Library, and the Nursing & Allied Health Database (NAHD) for literature, using the search strategies mentioned below. A final search was conducted on January 5, 2022, in the 3 literature databases as an update.

For PubMed, the following search was conducted: (clinical decision support systems[MeSH Terms]) AND (ontolog*[Title/abstract] OR rule*[Title/abstract]). For ACM Digital Library, the following search was conducted within the scope of the ACM Guide to Computing Literature: [[Publication Title: "clinical decision support*"] OR [Publication Title: cds*]] AND [[Publication Title: ontolog*] OR [Abstract: ontolog*] OR [Publication Title: rule*] OR [Abstract: rule*]]. For the NAHD, the following search was limited to peer-reviewed publications: mesh(clinical decision support) AND (ti(ontology)

OR ti(ontologies) OR ab(ontology) OR ab(ontologies) OR ti(rule) OR ti(rules) OR ab(rule) OR ab(rules)).

Figure 1. General workflow of the systematic literature review.



Inclusion and Exclusion Criteria

The inclusion criteria were as follows: text was written in English; full-text publication was available; ontologies were designed to be implemented or were already implemented in CDSSs, particularly related to CDSS rules; content included the granularity of CDSS rules; ontologies were designed to be integrated or were already integrated with health information systems (eg, EHRs), either in a production system or a prototype, with at least one architecture diagram, applied in clinical domains or designed for clinical domains to support health care providers; the publication was peer-reviewed; and details on the integration of CDSSs and EHRs were present for evaluation studies.

The exclusion criteria were as follows: only CDSS rules were included, regardless of the stage of the CDSS rule lifecycle (ie, development, identification, refinement, validation, evaluation, or implementation) or there was no mention of integration or ontologies; only ontologies were developed, evaluated, and validated, or there was no mention of integration or a CDSS; the system was designed without mentioning the granularity of CDSS rules or ontologies; and nonclinical decisions, such as administrative or management decisions (eg, supply chain management), were described.

General Workflow for Screening Papers

The first 100 papers were screened by all 3 authors (XJ, HM, and YG) independently. The first 100 retrieved papers were initially screened by 1 author (XJ) to draft initial inclusion and exclusion criteria. The inclusion and exclusion criteria were refined and adjusted by 2 authors (HM and YG) during the iterative screening, review, analysis, and discussions. Further, 2 authors (HM and YG) replicated the screening, and all 3 authors discussed and validated the results. The rest of the papers were then screened by at least 2 authors (XJ and HM, or XJ and YG) independently to determine inclusion. Disagreements were discussed and resolved via iterative rounds of group meetings.

The screening and manual review processes were conducted independently and approved by at least 2 authors. The literature was first screened based on titles, abstracts, and full-text publications when needed. The papers that were included were then manually coded to provide more content analysis and synthesized evidence. The final results were shared among all the authors. All disagreements were settled through group discussions.

Reviewing, Coding, Analyzing, and Synthesizing Processes

We followed grounded theory during the reviewing and manual coding of the included publications. One author (XJ) randomly selected 10 papers from the included 81 papers to start the coding (annotating) based on the focus of this literature review. ATLAS.ti 9 (desktop and web versions; ATLAS.ti Scientific Software Development GmbH), a qualitative data analytic tool, was used for coding. The coding results were discussed by 3 authors (XJ, HM, and YG). The discussion results formed the first draft of codes and code groups (Multimedia Appendix 1), that is, data items. Three coders (XJ, HM, and YG) then reviewed and coded the first 40 of the included papers using the initial principles and code groups, and added new codes and code groups when needed. Then, a second set of meetings was used to obtain consensus on updated principles and code groups. Refined codes and code groups were used to code the remaining papers. Every paper was coded by at least 2 coders independently. The coding results were then compared, and any discrepancies were resolved by group discussions. The code groups and codes were revised, consolidated, and updated during each discussion. Multimedia Appendix 2 presents the refined code groups and examples. Data items emerged during the review and were refined via discussions instead of predefinition before reviewing. Multimedia Appendix 3 lists all included papers.

After coding, the literature was analyzed and synthesized with a focus on several aspects, including CDSS application domains, CDSS mechanisms used in clinical settings, CDSS rule formats, authoring, management, and the roles of ontologies. The 3 authors worked together in an iterative process of analysis and synthetization. After obtaining consensus among all 3 authors, the results were then shared and discussed among all authors. Any concerns, confusions, or disagreements among the authors were resolved through iterative discussions. We followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 checklist [50] for reporting the systematic review with all relevant items (Multimedia Appendix 4 and Multimedia Appendix 5).

Results

Overview

By January 5, 2022, literature searches retrieved 1235 publications from 3 sources. After removing duplicates and examining according to the inclusion and exclusion criteria, 81

publications (Multimedia Appendix 3) were included in the final review and analysis [26,27,29,31-33,51-125]. Figure 2 depicts the literature search, screening, selection flow, and results. Figure 3 summarizes the main components covered by the literature review and the summary findings, and serves as an initial knowledge framework on CDSSs, CDSS rules, and ontology applications in CDSSs.

Figure 2. Flowchart of the literature search, screening, and selection. ACM: Association for Computing Machinery.

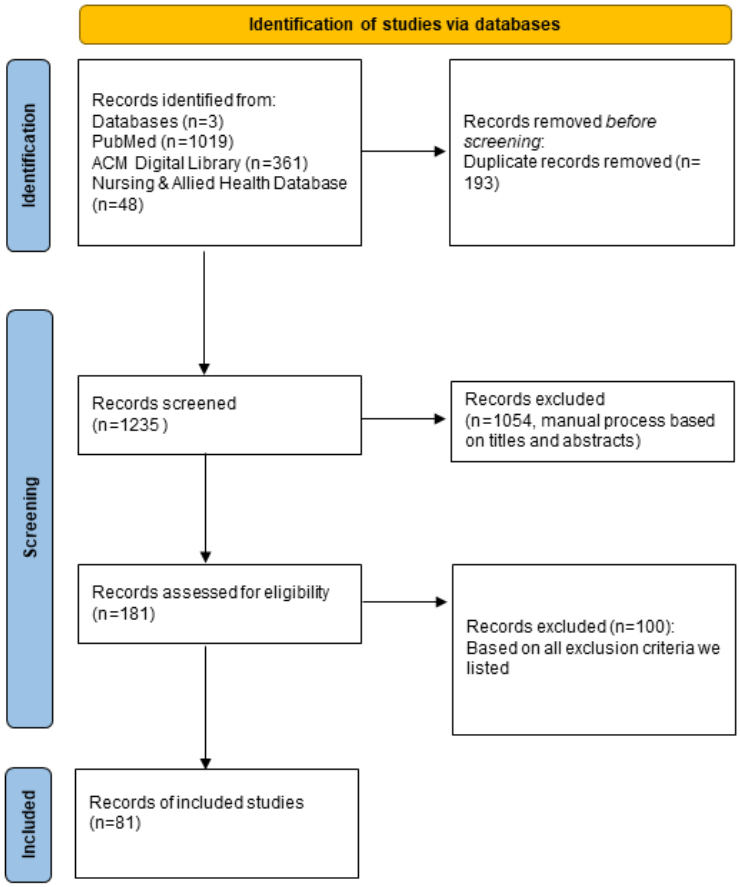
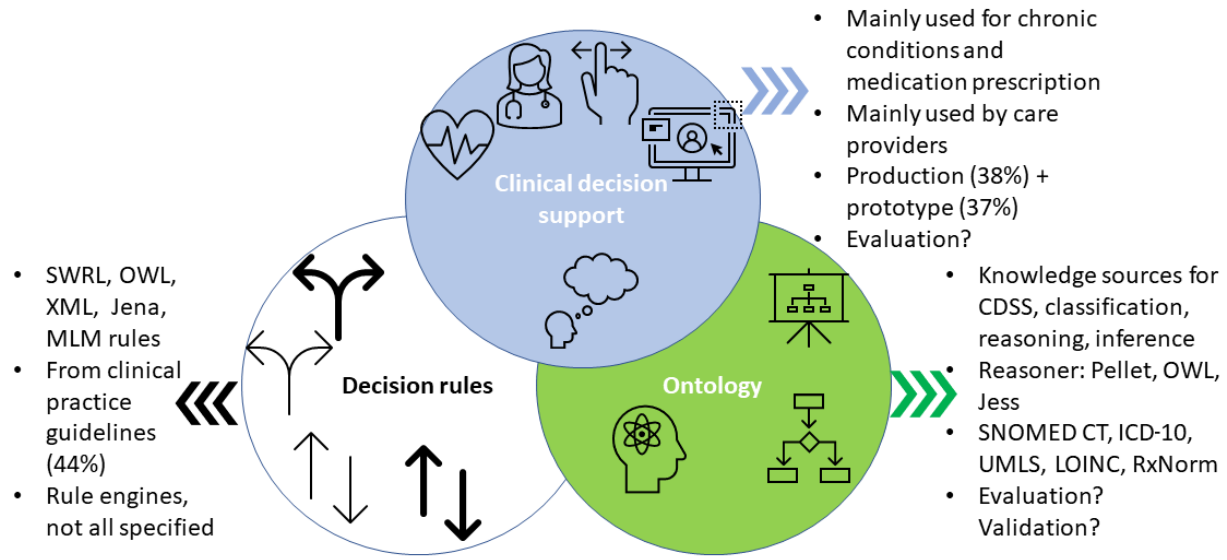


Figure 3. Initial knowledge framework on clinical decision support systems (CDSSs), CDSS rules, and ontology applications in CDSSs. ICD-10: International Statistical Classification of Diseases and Related Health Problems, 10th Revision; LOINC: Logical Observation Identifiers Names and Codes; MLM: medical logic module; OWL: Web Ontology Language; SNOMED CT: Systematized Nomenclature of Medicine-Clinical Terms; SWRL: Semantic Web Rule Language; UMLS: Unified Medical Language System; XML: Extensive Markup Language.



The majority of the publications (73/81, 90%) included in the review were from PubMed, a dominant source. After removing duplicates, the ACM Library added 8 new publications. After cleaning, discussion, and consolidation, 30 code groups and

221 final codes were used in ATLAS.ti (Multimedia Appendix 2). These codes and code groups guided our analysis and synthesis of the results. Multimedia Appendix 6 shows a word cloud image generated by ATLAS.ti that reflects the codes coded in the publications included.

PRISMA 2020 is designed to guide the reporting of outcome-oriented studies. Our systematic literature review focused on the design, development, and implementation of CDSSs, particularly related to CDSS rules and ontologies. Therefore, effect measures or certainty assessments were irrelevant items. We reported 19 categorical items (out of 27 categorical items, 26 items out of 42 items; Multimedia Appendix 4) for the full-text papers and 10 for the abstracts (out of 12 items; Multimedia Appendix 5).

Results Related to CDSS Characteristics

Over one-third (29/81, 36%) of CDSSs were designed and used for chronic condition management, prediction, or risk assessment, including but not limited to type 1 and 2 diabetes, hypertension, and asthma. Medication prescriptions (13/81, 16%), such as medication ordering, detection of adverse drug events, drug-drug interactions, and cancer care (8/81, 10%), were also significant application domains. Multimedia Appendix 7 illustrates the clinical domains of CDSSs within the included publications. Most CDSSs were designed for health care providers, but only 11% (9/81) were intended for patients. Most CDSSs provided recommendations, suggestions, alerts, or reminders. Among all the items in our comparison (Multimedia Appendix 8), EHR evaluation studies within the operational systems or prototypes exhibited the least complete information. Evaluations of CDSSs have been listed in multiple columns in Multimedia Appendix 8. Some CDSSs were implemented in production systems (31/81, 38%), whereas others were implemented in prototypes (30/81, 37%), which included experimental systems. Multimedia Appendix 8 summarizes the key features of CDSSs identified in the publications. In all tables, we adopted the original terms used in the corresponding papers. Some papers, for example, referred to “physicians” as CDSS users, whereas others referred to “clinicians” as CDSS users.

Results Related to CDSS Rules

Most CDSS rules were written in Web Ontology Language (OWL; 11/81, 14%), Extensive Markup Language (XML; 10/81, 12%), SWRL (9/81, 11%), Jena rules (5/81, 6%), and medical logic module (MLM; 3/81, 4%). Moreover, 2 publications [117,119] used N3 Language and 2 [90,117] used Natural Rule Language (NRL). Multimedia Appendix 9 presents 54 publications with more details on the CDSS rules, that is, publications that can fill out 3 or more cells (except for authors and publication year).

The most significant CDSS rule source is from clinical practice guidelines (36/81, 44%). Other sources of CDSS rules included domain expert input, publications (eg, textbooks and papers), multimedia sources, and internet resources. Data mining results were involved in CDSS rule sources [67,73]. CDSS rule authoring and editing tools were not routinely specified in the publications. Protégé [115] was the most prevalent tool to edit and author CDSS rules. Several publications also described developing authoring and editing tools [57,65,91].

There was a lack of technical details regarding rule engines, among which Jena (6/81, 7%), inference engine (6/81, 7%), Jess (4/81, 5%), JBoss (3/81, 4%), guideline engine (3/81, 4%), Drools (2/81, 3%), and Bayes (2/81, 3%) were frequently mentioned. Multimedia Appendix 9 summarizes how the CDSS rule (operation) works in a simplified manner. Many publications did not specify the working mechanism of CDSS rules within the EHR, electronic medical record (EMR), or hospital information system (HIS) context.

The majority of the publications did not appear to be focused on interoperability. Few papers that discussed interoperability (Multimedia Appendix 9) used HL7 CDA (Health Level 7, Clinical Document Architecture) or HL7 FHIR (Health Level 7, Fast Healthcare Interoperability Resources) standards. However, it is worth noting that such HL7 measures were not specifically designed for CDSS rules but rather for CDSS input and output.

Furthermore, some publications lacked necessary information for explaining the mechanisms of the systems, which can be critical barriers to reproducibility. Some publications lacked critical information, such as CDSS architecture diagrams; CDSS rule engines; CDSS rule languages; backend management methods for CDSS rules; and integration mechanisms among CDSS rules, ontologies, and EHR, EMR, or HIS systems.

Results Related to Ontologies

In the included publications, ontologies were primarily used as knowledge sources for CDSSs (32/81, 40%) to facilitate classification (7/81, 9%), reasoning, and inference (6/81, 7%; eg, identification recommendations or relationships). Moreover, ontologies were used to specify CDSS rules (12/81, 15%) or to provide general knowledge for the EMR or EHR systems. These 2 applications overlapped in some cases (19/81, 24%; ie, the ontologies were used to provide specified CDSS rules and general knowledge).

In the included publications, the terms “reasoner” and “rule engines” were used interchangeably. *Reasoner*, in our opinion, refers to the inference for a consistency check or classification for an ontology. A reasoner can be part of an ontology tool or can be external. For CDSSs, a *rule engine* is the mechanism that generates or provides recommendations by incorporating a patient’s data, contextual information, and medical knowledge (typically from an ontology or knowledge base). However, we kept the authors’ choice of terms in tables without modification. Among the included publications, the most common reasoners were Pellet (11/81, 14%), Jena (4/81, 5%), OWL reasoner (3/81, 4%), Jess (2/81, 3%), and the Euler/EYE inference engine (2/81, 3%).

The content and code systems used to represent the content should be included as ontology sources. The content could come from a popular textbook or a clinical practice guideline. The content can be coded in a specific code system, such as SNOMED CT (Systematized Nomenclature of Medicine-Clinical Terms). Multimedia Appendix 10 includes code systems that served as ontology sources. The most often

used coding systems among the included publications were SNOMED CT (9/81, 11%), the International Statistical Classification of Diseases and Related Health Problems, 10th Revision (ICD-10; 4/81, 5%), Unified Medical Language System (UMLS; 4/81, 5%), Logical Observation Identifiers Names and Codes (LOINC; 3/81, 4%), and RxNorm (2/81, 3%).

Incomplete information for ontology validation is a common issue shown in the literature. Approximately 20 publications mentioned some validation, including validation or evaluation by domain experts (20/81, 25%). Some ontologies were authored by domain experts [55,63]. Multimedia Appendix 10 provides more information on the roles of ontologies in publications, including publications with 3 or more cells (except for authors and publication year; n=36).

Discussion

Summary of the Results

Although ontologies contribute to the content of CDSS rules and have the potential to facilitate interoperable CDSS rules, our systematic review showed that reusing and sharing of CDSS rules have not been achieved. CDSSs have a wide range of clinical application domains, primarily for health care providers, such as chronic condition management, medication ordering, and cancer care. CDSS rules are primarily based on clinical practice guidelines.

Although reusing and sharing CDSS artifacts are well-recognized challenges [1,109], reusability, customization, and shareability of CDSS rules are not yet a common focus, even in publications focusing on CDSS rule editing [43,126,127]. These are important topics to cover in a literature review. Marco-Ruiz et al [109] demonstrated how to use CDSS artifacts in the Linked Data framework [128] by leveraging Semantic Web technologies, particularly ontology. However, that work was at a higher level, describing concepts without tangible tools implemented in clinical practice. To fill this gap, one approach is to build an upper-level CDSS ontology [129] to encourage the reuse of CDSS rules and demonstrate the potential of ontologies. Our effort is in alignment with their vision, as well as other efforts in reusing and sharing CDSS artifacts [1,109].

Ontologies were not at the center of any early examples of CDSSs [6-12]. An early demonstration of using medical terminology in CDSSs was the adoption of Current Medical Information and Terminology (CMIT) in a diagnostic engine [130,131]. Even under our “loose use of ontology” during our systematic literature search, there was no case in which ontology played a central role in sharing CDSS rules, particularly for rule management and maintenance.

Over the years, CDSSs have been successfully applied in clinical care. Unfortunately, CDSS rules are not yet portable. Making CDSS rules more portable is therefore significant work that could be leveraged by ontologies, and our systematic literature review brings us one step closer to that goal. Marco-Ruiz et al also conducted a very relevant systematic literature review. However, their focus was on the interoperability mechanisms used in CDSSs [132,133]. According to the results of their systematic literature review, 32% of the included papers used ontologies and 46% used standard terminologies. The findings related to ontologies are similar [132,133] to those of our paper. However, we presented a more detailed and thorough analysis of these technologies used in CDSS rules. Nevertheless, both papers concluded that complete CDSS interoperability is not a reality. Thus, additional efforts are required to achieve interoperable and reusable CDSS artifacts, such as CDSS rules.

Interpretation of the Results

Rule engines, which execute rules, patient data, and context information to produce a result, such as an alert or a recommendation, are critical components of CDSSs [1]. Jess, a rule engine and development environment in Java [134], was frequently mentioned in the included publications as a tool for developing rule-based CDSSs. SWRL rules can be converted to Jess rules in the popular tool Protégé, using a plug-in application programming interface (API) SWRLJessTab. Jess rules can be used by the Jess rule engine, which is widely used in rule-based expert systems [134]. In addition to Jess, Jena and Drools were used frequently in the publications included. Jena is a Java API that supports rule-based inference and makes use of resource description framework (RDF) graphs [135]. Jena.java API is a popular framework for managing RDF/OWL descriptions and can handle OWL models [96]. Drools is a business rule management system that includes a rule engine [136]. Drools also has the SWRL API that supports SWRL and Semantic Query-Enhanced Web Rule Language (SQWRL). SWRL can be queried by SQWRL.

Reasoning via a reasoner is a critical characteristic of many ontologies, even though the current reasoning is still in first-degree logic. Reasoning can be used for the following 3 main functions: consistency check, classification, and realization [137]. Several publications specified the classification roles of the ontologies and reasoners (Multimedia Appendix 10). The Manchester University OWL group has curated an updated list of OWL reasoners [137]. Parsia et al [137] compiled and compared the current OWL reasoners and their performances via the competition report. Both Pellet and Jena are popular reasoners (Multimedia Appendix 10), and other reasoners include FaCT++ [98], Z3 Solver reasoner [105], Euler/EYE inference engine [117,119], OWL Horst [109], and OWL Cerebra [63] among the included publications. Among these reasoners, Pellet [138] is Java based, and it can work on SWRL rules and ontologies written in OWL2. SWRL was initially designed as a rule language for Semantic Web technologies [139]. A user needs the rule language and an editor (eg, Protégé SWRL tab) to write, revise, and query the rules. SWRL can be queried by SQWRL (a query language for OWL) or SPARQL (SPARQL Protocol and RDF query language). Reasoners can then be used to conduct reasoning based on the rules and facts defined in the ontology or knowledge base. Protégé-OWL [140] provides an editor for SWRL rules. Protégé SWRL editor is another example.

This review has demonstrated unique insights about CDSS rules, ontologies, and ontology applications, particularly in CDSS rule management and maintenance, and has presented several distinct characteristics that complement the existing literature.

An earlier review [40] focused on clinical decision-making in forming ontologies to support complex cognitive processes and reasoning processes comparing evaluation metrics but did not cover the implementation of EHR, EMR, or HIS systems and the mechanisms of these characteristics.

Significance of the Work

Our systematic review demonstrated the state-of-the-art applications of ontologies in CDSS rules. These applications have a lot of potential for reusing and sharing CDSS artifacts. However, none of the existing papers elaborated or demonstrated how ontologies enable portable CDSS rules. Although some authors recognized this benefit [1,43,109], none have conducted a systematic review. Our literature review thoroughly examined the topic, outlined the current frontlines on CDSS rules and ontology uses in CDSSs, established the knowledge framework, and compiled a comprehensive collection of relevant publications that can inform future efforts to design or improve CDSSs. This systematic review focused on the mechanisms of CDSSs in clinical practices or prototypes, CDSS rules, and ontology roles in CDSSs. The detailed information provided in each included publication (Multimedia Appendix 8, Multimedia Appendix 9, and Multimedia Appendix 10) about the reasoners, rule engines, ontologies, and CDSS rule formats used provided valuable references for designing or improving systems. The side-by-side comparison of publications (Multimedia Appendix 8, Multimedia Appendix 9, and Multimedia Appendix 10) also provided structured guidance for preparing future designs and publications or teaching references on the topics in tangible ways.

Missing Information in the Publications and Our Recommendations

Inconsistent or missing information about CDSS rule languages, CDSS rule engines, and CDSS evaluation details was identified. In CDSS evaluation, there was commonly no information about how the evaluation was conducted or who performed the evaluation. There were also inconsistencies in technical details related to ontology purposes, reasoners, connection mechanisms, or communications between CDSSs and EHR, EMR, or HIS systems. Inconsistent or missing information hampered reproducibility and further improvement of published work. We are obviously not the only group that has identified missing critical information as a problem in technical papers on similar topics [141].

Another missing piece is the evaluation and validation of ontologies or knowledge bases. Only 25% of publications mentioned that domain experts conducted evaluation or validation. A formal assessment or validation is critical to ensure the validity of the results from automated processes for some ontologies (or knowledge bases) derived from other automatic methods (eg, machine learning algorithms). Testing has not been conducted consistently across the publications. Some ontologies were authored by domain experts, which provides greater validity than those involving nondomain experts while constructing ontologies.

Thus, it is recommended that authors include essential technical details in publications. These technical details include CDSS application domains, intended CDSS users, CDSS notification types, CDSS evaluations (what, how, and by whom), CDSS rule sources, CDSS rule languages, CDSS rule engines, CDSS operation mechanisms, ontology use purposes, ontology sources (both content and code systems), ontology validation, reasoners, and connection or communication mechanisms between CDSSs and EMR, EHR, or HIS systems. Authors are highly encouraged to include such details to help readers reference, compare, and increase the reproducibility of the reported work.

Limitations

Our review has limitations. Non-English publications or full-text unavailable publications were not included. Publications that focused only on CDSS rules [43,126,127] were also excluded. Moreover, publications without specifying an ontology component were excluded, although such publications had a similar focus to one aspect of our systematic review. We also noticed that most of the publications on CDSS rule authoring and managing tools were from Partners HealthCare/Harvard Medical School. The strengths of Partners HealthCare/Harvard Medical School were shown. On the other hand, a lack of broad adoption, implementation, or publication of such topics was shown.

When “CDSS” is not specified as a keyword, the search results may exclude publications. For example, our 2 previous papers [142,143] were not found via the search strategy because “CDSS” was not used as a keyword, although the content was undoubtedly within the scope of this review. This challenge is common to how our current literature databases are organized and how we conduct a literature search. Even with MeSH (Medical Subject Headings; the controlled vocabulary for PubMed), publications can still be missed without using commonly recognized keywords. This challenge could be minimized and mitigated by carefully developing an exhaustive list of keywords to maximize the possibilities found during a literature search in the future.

Conclusions

The reuse, management, and maintenance of CDSS rules are critical yet challenging for their clinical application. Although ontologies have been used to contribute to the content of CDSS rules, they have not been used to facilitate CDSS rule reuse and sharing. Building a CDSS ontology, which could be the first tangible step, requires bridging high-level visions and operational efforts. Semantic interoperability remains a major challenge that must be overcome to achieve reuse of CDSS artifacts, including CDSS rules. The realization of semantic interoperability will not only allow for the reuse of CDSS artifacts, which are resource intensive to develop and maintain, but also provide practical insights to achieve interoperable patient records. This has been a long-lost aspect, and health care providers will be able to access patients’ complete records to provide safer and higher quality care every time to every patient. We believe that making CDSS rules interoperable can provide insightful guidance for interoperable patient records.

Incomplete technical details on CDSS rules and ontologies presented in publications should be addressed in future publications by including more detailed information about

architectural diagrams; the mechanisms of connection among ontologies, CDSS rules, and EHR, EMR, or HIS systems; CDSS rule languages; reasoners; rule engines; the validation or authorization of ontologies and CDSS rules; the purposes of ontologies; ontology sources; and the management and maintenance of CDSS rules. Such information can help researchers to optimize design and development while also increasing reproducibility. Finally, the knowledge framework and the summarization of included publications are expected to guide future CDSS improvements and innovations, CDSS rules, and the integration and communication of CDSSs with EHR, EMR, or HIS systems.

Acknowledgments

The authors thank professors James J Cimino, Dean F Sittig, and Adam Wright for their valuable feedback and suggestions on the manuscript. This work was supported by the National Institute of General Medical Sciences of the National Institutes of Health (under award number R01GM138589 and partially under award number P20GM121342). The content is solely the authors' responsibility and does not necessarily represent the official views of the National Institutes of Health.

Data Availability

A request for additional data in addition to the results and appendices of this manuscript can be made to the corresponding author, and the final decision on data release will be made on a case-by-case basis, as appropriate.

Authors' Contributions

XJ, HM, and YG designed the study, conducted the review and coding, drafted the first version, analyzed and interpreted the results, and revised the manuscript significantly. PB, DR, TL, CN, AF, LR, NH, and RG participated in the design of the study, analyzed and interpreted the results, and revised the manuscript significantly.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Initial draft of codes and code groups used in reviewing/coding.
[PDF File (Adobe PDF File), 195 KB - medinform_v11i1e43053_app1.pdf]

Multimedia Appendix 2
Refined version of codes and code groups used in reviewing/coding.
[PDF File (Adobe PDF File), 215 KB - medinform_v11i1e43053_app2.pdf]

Multimedia Appendix 3
List of papers included in this systematic literature review (n=81).
[PDF File (Adobe PDF File), 222 KB - medinform_v11i1e43053_app3.pdf]

Multimedia Appendix 4
PRISMA 2020 item checklist reported in our systematic literature review.
[PDF File (Adobe PDF File), 155 KB - medinform_v11i1e43053_app4.pdf]

Multimedia Appendix 5
PRISMA 2020 for Abstracts checklist reported in our systematic literature review.
[PDF File (Adobe PDF File), 37 KB - medinform_v11i1e43053_app5.pdf]

Multimedia Appendix 6
Word cloud generated from ATLAS.ti based on codes in the included publications .
[PNG File , 353 KB - medinform_v11i1e43053_app6.png]

Multimedia Appendix 7
Bar chart generated from ATLAS.ti showing the clinical domains of clinical decision support systems in included publications.
[PNG File , 122 KB - medinform_v11i1e43053_app7.png]

Multimedia Appendix 8

Basic clinical decision support system profiles in included publications (n=81).
[PDF File (Adobe PDF File), 198 KB - medinform_v11i1e43053_app8.pdf]

Multimedia Appendix 9
Comparison of clinical decision support system rule characteristics in included publications (n=54).
[PDF File (Adobe PDF File), 182 KB - medinform_v11i1e43053_app9.pdf]

Multimedia Appendix 10
Comparison of ontology roles in included publications (n=36).
[PDF File (Adobe PDF File), 151 KB - medinform_v11i1e43053_app10.pdf]

References

1. Greenes RA. Clinical Decision Support: The Road to Broad Adoption. Cambridge, MA: Academic Press; 2014.
2. Berner ES. Clinical decision support systems: State of the Art. AHRQ. Rockville, MD: Agency for Healthcare Research and Quality; 2009. URL: https://digital.ahrq.gov/sites/default/files/docs/biblio/09-0069-EF_1.pdf [accessed 2022-12-21]
3. Sittig DF, Wright A, Meltzer S, Simonaitis L, Evans RS, Nichol WP, et al. Comparison of clinical knowledge management capabilities of commercially-available and leading internally-developed electronic health records. BMC Med Inform Decis Mak 2011 Feb 17;11:13 [FREE Full text] [doi: 10.1186/1472-6947-11-13] [Medline: 21329520]
4. Sittig DF, Wright A, Simonaitis L, Carpenter JD, Allen GO, Doebbeling BN, et al. The state of the art in clinical knowledge management: an inventory of tools and techniques. Int J Med Inform 2010 Jan;79(1):44-57 [FREE Full text] [doi: 10.1016/j.ijmedinf.2009.09.003] [Medline: 19828364]
5. Moja L, Polo Friz H, Capobussi M, Kwag K, Banzi R, Ruggiero F, et al. Effectiveness of a Hospital-Based Computerized Decision Support System on Clinician Recommendations and Patient Outcomes: A Randomized Clinical Trial. JAMA Netw Open 2019 Dec 02;2(12):e1917094 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.17094] [Medline: 31825499]
6. Pople HE, Myers JD, Miller RA. DIALOG: a model of diagnostic logic for internal medicine. In: IJCAI'75: Proceedings of the 4th International Joint Conference on Artificial Intelligence - Volume 1. 1975 Presented at: 4th International Joint Conference on Artificial Intelligence; September 3-8, 1975; Tbilisi, Georgia, USSR p. 848-855. [doi: 10.5555/1624626.1624759]
7. Miller RA. A history of the INTERNIST-1 and Quick Medical Reference (QMR) computer-assisted diagnosis projects, with lessons learned. Yearb Med Inform 2010:121-136. [Medline: 20938584]
8. Miller RA, McNeil MA, Challinor SM, Masarie FE, Myers JD. The INTERNIST-1/QUICK MEDICAL REFERENCE project--status report. West J Med 1986 Dec;145(6):816-822 [FREE Full text] [Medline: 3544509]
9. Miller RA, Pople HE, Myers JD. Internist-1, an experimental computer-based diagnostic consultant for general internal medicine. N Engl J Med 1982 Aug 19;307(8):468-476. [doi: 10.1056/NEJM198208193070803] [Medline: 7048091]
10. Bergeron B. Iliad: a diagnostic consultant and patient simulator. MD Comput 1991;8(1):46-53. [Medline: 1822085]
11. Warner HR. Iliad: diagnostic tools for general medicine. Interview by Bill W. Childs. Healthc Inform 1990 Apr;7(4):38. [Medline: 10120646]
12. Warner HR, Bouhaddou O. Innovation review: Iliad--a medical diagnostic support program. Top Health Inf Manage 1994 May;14(4):51-58. [Medline: 10134761]
13. Lobach D, Sanders G, Bright T, Wong A, Dhurjati R, Bristow E, et al. Enabling health care decisionmaking through clinical decision support and knowledge management. Evid Rep Technol Assess (Full Rep) 2012 Apr(203):1-784. [Medline: 23126650]
14. Kawamoto K, Houlihan CA, Balas EA, Lobach DF. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. BMJ 2005 Apr 02;330(7494):765 [FREE Full text] [doi: 10.1136/bmj.38398.500764.8F] [Medline: 15767266]
15. O'Connor PJ, Sperl-Hillen JM, Rush WA, Johnson PE, Amundson GH, Asche SE, et al. Impact of electronic health record clinical decision support on diabetes care: a randomized trial. Ann Fam Med 2011;9(1):12-21 [FREE Full text] [doi: 10.1370/afm.1196] [Medline: 21242556]
16. McDonald CJ. Protocol-based computer reminders, the quality of care and the non-perfectability of man. N Engl J Med 1976 Dec 09;295(24):1351-1355. [doi: 10.1056/NEJM197612092952405] [Medline: 988482]
17. Middleton B, Sittig DF, Wright A. Clinical Decision Support: a 25 Year Retrospective and a 25 Year Vision. Yearb Med Inform 2016 Aug 02;Suppl 1(Suppl 1):S103-S116 [FREE Full text] [doi: 10.15265/IYS-2016-s034] [Medline: 27488402]
18. Wright A, Ash JS, Aaron S, Ai A, Hickman TT, Wiesen JF, et al. Best practices for preventing malfunctions in rule-based clinical decision support alerts and reminders: Results of a Delphi study. Int J Med Inform 2018 Oct;118:78-85 [FREE Full text] [doi: 10.1016/j.ijmedinf.2018.08.001] [Medline: 30153926]
19. Horsky J, Schiff GD, Johnston D, Mercincavage L, Bell D, Middleton B. Interface design principles for usable decision support: a targeted review of best practices for clinical prescribing interventions. J Biomed Inform 2012 Dec;45(6):1202-1216 [FREE Full text] [doi: 10.1016/j.jbi.2012.09.002] [Medline: 22995208]

20. Wu R, Peters W, Morgan MW. The next generation of clinical decision support: linking evidence to best practice. J Healthc Inf Manag 2002;16(4):50-55. [Medline: 12365300]
21. Fox J, Gutenstein M, Khan O, South M, Thomson R. OpenClinical.net: A platform for creating and sharing knowledge and promoting best practice in healthcare. Computers in Industry 2015 Jan;66:63-72. [doi: 10.1016/j.compind.2014.10.001]
22. Teich JM, Osheroff JA, Pifer EA, Sittig DF, Jenders RA, CDS Expert Review Panel. Clinical decision support in electronic prescribing: recommendations and an action plan: report of the joint clinical decision support workgroup. J Am Med Inform Assoc 2005;12(4):365-376 [FREE Full text] [doi: 10.1197/jamia.M1822] [Medline: 15802474]
23. Sittig DF, Wright A, Osheroff JA, Middleton B, Teich JM, Ash JS, et al. Grand challenges in clinical decision support. J Biomed Inform 2008 Apr;41(2):387-392 [FREE Full text] [doi: 10.1016/j.jbi.2007.09.003] [Medline: 18029232]
24. Cimino JJ. An integrated approach to computer-based decision support at the point of care. Trans Am Clin Climatol Assoc 2007;118:273-288 [FREE Full text] [Medline: 18528510]
25. Jing X, Himawan L, Law T. Availability and usage of clinical decision support systems (CDSSs) in office-based primary care settings in the USA. BMJ Health Care Inform 2019 Dec 08;26(1):e100015 [FREE Full text] [doi: 10.1136/bmjhci-2019-100015] [Medline: 31818828]
26. Borbolla D, Otero C, Lobach DF, Kawamoto K, Gomez Saldaño AM, Staccia G, et al. Implementation of a clinical decision support system using a service model: results of a feasibility study. Stud Health Technol Inform 2010;160(Pt 2):816-820. [Medline: 20841799]
27. Goldberg HS, Paterno MD, Rocha BH, Schaeffer M, Wright A, Erickson JL, et al. A highly scalable, interoperable clinical decision support service. J Am Med Inform Assoc 2014 Feb 01;21(e1):e55-e62 [FREE Full text] [doi: 10.1136/amiajnl-2013-001990] [Medline: 23828174]
28. Miller PL, Frawley SJ, Sayward FG, Yasnoff WA, Duncan L, Fleming DW. IMM/Serve: a rule-based program for childhood immunization. Proc AMIA Annu Fall Symp 1996:184-188 [FREE Full text] [Medline: 8947653]
29. Nguyen B, Reese T, Decker S, Malone D, Boyce RD, Beyan O. Implementation of Clinical Decision Support Services to Detect Potential Drug-Drug Interaction Using Clinical Quality Language. Stud Health Technol Inform 2019 Aug 21;264:724-728. [doi: 10.3233/SHTI190318] [Medline: 31438019]
30. Linder J, Schnipper JL, Volk LA, Tsurikova R, Palchuk M, Olsha-Yehiav M, et al. Clinical decision support to improve antibiotic prescribing for acute respiratory infections: results of a pilot study. AMIA Annu Symp Proc 2007 Oct 11;2007:468-472 [FREE Full text] [Medline: 18693880]
31. Ray HN, Boxwala AA, Anantraman V, Ohno-Machado L. Providing context-sensitive decision-support based on WHO guidelines. Proc AMIA Symp 2002:637-641 [FREE Full text] [Medline: 12463901]
32. Barth C, Tobman M, Nätscher C, Sussmann H, Horsch A. Fusing a systematic and a case-based repository for medical decision support. Stud Health Technol Inform 2003;95:560-564. [Medline: 14664046]
33. Haug PJ, Ferraro JP, Holmen J, Wu X, Mynam K, Ebert M, et al. An ontology-driven, diagnostic modeling system. J Am Med Inform Assoc 2013 Jun 01;20(e1):e102-e110 [FREE Full text] [doi: 10.1136/amiajnl-2012-001376] [Medline: 23523876]
34. Del Fiol G, Curtis C, Cimino JJ, Iskander A, Kalluri ASD, Jing X, et al. Disseminating context-specific access to online knowledge resources within electronic health record systems. Stud Health Technol Inform 2013;192:672-676 [FREE Full text] [Medline: 23920641]
35. Del Fiol G, Williams MS, Maram N, Rocha RA, Wood GM, Mitchell JA. Integrating genetic information resources with an EHR. AMIA Annu Symp Proc 2006;2006:904 [FREE Full text] [Medline: 17238523]
36. Cimino JJ, Overby CL, Devine EB, Hulse NC, Jing X, Maviglia SM, et al. Practical choices for infobutton customization: experience from four sites. AMIA Annu Symp Proc 2013;2013:236-245 [FREE Full text] [Medline: 24551334]
37. Del Fiol G, Haug PJ, Cimino JJ, Narus SP, Norlin C, Mitchell JA. Effectiveness of topic-specific infobuttons: a randomized controlled trial. J Am Med Inform Assoc 2008;15(6):752-759 [FREE Full text] [doi: 10.1197/jamia.M2725] [Medline: 18755999]
38. Cimino JJ. Infobuttons: anticipatory passive decision support. AMIA Annu Symp Proc 2008 Nov 06:1203-1204. [Medline: 18998777]
39. Patkar V, Acosta D, Davidson T, Jones A, Fox J, Keshtgar M. Using computerised decision support to improve compliance of cancer multidisciplinary meetings with evidence-based guidance. BMJ Open 2012;2(3):e000439 [FREE Full text] [doi: 10.1136/bmjopen-2011-000439] [Medline: 22734113]
40. Sutton DR, Fox J. The Syntax and Semantics of the PRO Guideline Modeling Language. J Am Med Inform Assoc 2003 Sep 01;10(5):433-443. [doi: 10.1197/jamia.m1264]
41. Del Fiol G, Kohlmann W, Bradshaw RL, Weir CR, Flynn M, Hess R, et al. Standards-Based Clinical Decision Support Platform to Manage Patients Who Meet Guideline-Based Criteria for Genetic Evaluation of Familial Cancer. JCO Clinical Cancer Informatics 2020 Nov(4):1-9. [doi: 10.1200/cci.19.00120]
42. Hripcsak G. Writing Arden Syntax Medical Logic Modules. Comput Biol Med 1994 Sep;24(5):331-363. [doi: 10.1016/0010-4825(94)90002-7] [Medline: 7705066]
43. Zhou L, Karipineni N, Lewis J, Maviglia SM, Fairbanks A, Hongsermeier T, et al. A study of diverse clinical decision support rule authoring environments and requirements for integration. BMC Med Inform Decis Mak 2012 Nov 12;12(1):128 [FREE Full text] [doi: 10.1186/1472-6947-12-128] [Medline: 23145874]

44. Davies J, Fensel D, van Harmelen F. Towards the Semantic Web: Ontology-driven Knowledge Management. Hoboken, NJ: John Wiley & Sons; 2002.
45. What is an Ontology? Stanford Knowledge Systems, AI Laboratory. URL: http://www-ksl.stanford.edu/kst/what-is-an-ontology.html [accessed 2022-12-21]
46. Musen MA. Dimensions of knowledge sharing and reuse. Comput Biomed Res 1992 Oct;25(5):435-467. [doi: 10.1016/0010-4809(92)90003-s] [Medline: 1395522]
47. Whetzel PL, Noy NF, Shah NH, Alexander PR, Nyulas C, Tudorache T, et al. BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. Nucleic Acids Res 2011 Jul;39(Web Server issue):W541-W545 [FREE Full text] [doi: 10.1093/nar/gkr469] [Medline: 21672956]
48. Dissanayake PI, Colicchio TK, Cimino JJ. Using clinical reasoning ontologies to make smarter clinical decision support systems: a systematic review and data synthesis. J Am Med Inform Assoc 2020 Jan 01;27(1):159-174 [FREE Full text] [doi: 10.1093/jamia/ocz169] [Medline: 31592534]
49. Samwald M, Stenzhorn H, Dumontier M, Marshall MS, Luciano J, Adlassnig K. Towards an interoperable information infrastructure providing decision support for genomic medicine. Stud Health Technol Inform 2011;169:165-169. [Medline: 21893735]
50. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. J Clin Epidemiol 2021 Jun;134:178-189 [FREE Full text] [doi: 10.1016/j.jclinepi.2021.03.001] [Medline: 33789819]
51. Manjarrés Riesco A, Martı´nez Tomás R, Mira Mira J. A customisable framework for the assessment of therapies in the solution of therapy decision tasks. Artificial Intelligence in Medicine 2000 Jan;18(1):57-82. [doi: 10.1016/s0933-3657(99)00029-9]
52. Abidi SS, Manickam S. Transforming XML-based electronic patient records for use in medical case based reasoning systems. Stud Health Technol Inform 2000;77:709-713. [Medline: 11187645]
53. De Clercq PA, Blom JA, Hasman A, Korsten HH. GASTON: an architecture for the acquisition and execution of clinical guideline-application tasks. Med Inform Internet Med 2000;25(4):247-263. [doi: 10.1080/146392300455558] [Medline: 11198187]
54. Payne TH, Savarino J, Marshall R, Hoey CT. Use of a clinical event monitor to prevent and detect medication errors. Proc AMIA Symp 2000:640-644 [FREE Full text] [Medline: 11079962]
55. Achour SL, Dojat M, Rieux C, Bierling P, Lepage E. A UMLS-based knowledge acquisition tool for rule-based clinical decision support system development. J Am Med Inform Assoc 2001;8(4):351-360 [FREE Full text] [doi: 10.1136/jamia.2001.0080351] [Medline: 11418542]
56. Séroussi B, Bouaud J, Dréau H, Falcoff H, Riou C, Joubert M, et al. ASTI: a guideline-based drug-ordering system for primary care. Stud Health Technol Inform 2001;84(Pt 1):528-532. [Medline: 11604796]
57. Karadimas HC, Chailloleau C, Hemery F, Simonnet J, Lepage E. Arden/J: an architecture for MLM execution on the Java platform. J Am Med Inform Assoc 2002;9(4):359-368 [FREE Full text] [doi: 10.1197/jamia.m0985] [Medline: 12087117]
58. Das AK, Musen MA. SYNCHRONUS: a reusable software module for temporal integration. Proc AMIA Symp 2002:195-199 [FREE Full text] [Medline: 12463814]
59. Poon EG, Wang SJ, Gandhi TK, Bates DW, Kuperman GJ. Design and implementation of a comprehensive outpatient Results Manager. J Biomed Inform 2003;36(1-2):80-91 [FREE Full text] [doi: 10.1016/s1532-0464(03)00061-3] [Medline: 14552849]
60. Liaw S, Sulaiman N, Pearce C, Sims J, Hill K, Grain H, et al. Falls Prevention within the Australian General Practice Data Model: Methodology, Information Model, and Terminology Issues. J Am Med Inform Assoc 2003 Sep 01;10(5):425-432. [doi: 10.1197/jamia.m1281]
61. Greenes RA, Sordo M, Zaccagnini D, Meyer M, Kuperman GJ. Design of a standards-based external rules engine for decision support in a variety of application contexts: report of a feasibility study at Partners HealthCare System. Stud Health Technol Inform 2004;107(Pt 1):611-615. [Medline: 15360885]
62. Ebrahiminia V, Riou C, Seroussi B, Bouaud J, Dubois S, Falcoff H, et al. Design of a decision support system for chronic diseases coupling generic therapeutic algorithms with guideline-based specific rules. Stud Health Technol Inform 2006;124:483-488. [Medline: 17108565]
63. Kashyap V, Morales A, Hongsermeier T. On implementing clinical decision support: achieving scalability and maintainability by combining business rules and ontologies. AMIA Annu Symp Proc 2006;2006:414-418 [FREE Full text] [Medline: 17238374]
64. Verlaene K, Joosen W, Verbaeten P. Arriclides: An Architecture Integrating Clinical Decision Support Models. 2007 Presented at: 40th Annual Hawaii International Conference on System Sciences (HICSS'07); January 3-6, 2007; Waikoloa, HI, USA. [doi: 10.1109/HICSS.2007.87]
65. Abidi SR. Ontology-Based Modeling of Breast Cancer Follow-up Clinical Practice Guideline for Providing Clinical Decision Support. 2007 Presented at: Twentieth IEEE International Symposium on Computer-Based Medical Systems (CBMS'07); June 20-22, 2007; Maribor, Slovenia. [doi: 10.1109/CBMS.2007.80]

66. Jannin P, Morandi X. Surgical models for computer-assisted neurosurgery. Neuroimage 2007 Sep 01;37(3):783-791 [FREE Full text] [doi: 10.1016/j.neuroimage.2007.05.034] [Medline: 17613249]
67. Stacey M, McGregor C, Tracy M. An architecture for multi-dimensional temporal abstraction and its application to support neonatal intensive care. Annu Int Conf IEEE Eng Med Biol Soc 2007;2007:3752-3756. [doi: 10.1109/IEMBS.2007.4353148] [Medline: 18002814]
68. Papageorgiou E, Stylios C, Groumpos P. Novel Architecture for supporting medical decision making of different data types based on Fuzzy Cognitive Map Framework. Annu Int Conf IEEE Eng Med Biol Soc 2007;2007:1192-1195. [doi: 10.1109/IEMBS.2007.4352510] [Medline: 18002176]
69. Cornalba C, Bellazzi R, Bellazzi R. Building a Normative Decision Support System for Clinical and Operational Risk Management in Hemodialysis. IEEE Trans. Inform. Technol. Biomed 2008 Sep;12(5):678-686. [doi: 10.1109/titb.2008.920781]
70. Farion K, Michalowski W, Wilk S, O'Sullivan D, Rubin S, Weiss D. Clinical Decision Support System for Point of Care Use. Methods Inf Med 2018 Jan 17;48(04):381-390. [doi: 10.3414/me0574]
71. Dao TT, Marin F, Ho Ba Tho MC. Clinical validated Computer-Aided Decision System to the clubfeet deformities. Annu Int Conf IEEE Eng Med Biol Soc 2009;2009:6230-6233. [doi: 10.1109/IEMBS.2009.5334650] [Medline: 19965086]
72. Zhou Q. A Clinical Decision Support System for Metabolism Synthesis. 2009 Presented at: International Conference on Computational Intelligence and Natural Computing; June 6-7, 2009; Wuhan, China. [doi: 10.1109/CINC.2009.89]
73. Carenini M, ReMINE Consortium. ReMINE: an ontology-based risk management platform. Stud Health Technol Inform 2009;148:32-42. [Medline: 19745233]
74. Lee J, Kim J, Cho I, Kim Y. Integration of workflow and rule engines for clinical decision support services. Stud Health Technol Inform 2010;160(Pt 2):811-815. [Medline: 20841798]
75. Ongenae F, Dhaene T, Turck F, Benoit D, Decruyenaere J. Design of a probabilistic ontology-based clinical decision support system for classifying temporal patterns in the ICU: A sepsis case study. 2010 Presented at: 23rd International Symposium on Computer-Based Medical Systems (CBMS); October 12-15, 2010; Bentley, WA, Australia. [doi: 10.1109/CBMS.2010.6042676]
76. Basilakis J, Lovell NH, Redmond SJ, Celler BG. Design of a decision-support architecture for management of remotely monitored patients. IEEE Trans Inf Technol Biomed 2010 Sep;14(5):1216-1226. [doi: 10.1109/TITB.2010.2055881] [Medline: 20615815]
77. Wilk S, Michalowski W, Farion K, Sayyad Shirabad J. MET3-AE system to support management of pediatric asthma exacerbation in the emergency department. Stud Health Technol Inform 2010;160(Pt 2):841-845. [Medline: 20841804]
78. Bouamrane M, Rector A, Hurrell M. Using OWL ontologies for adaptive patient information modelling and preoperative clinical decision support. Knowl Inf Syst 2010 Oct 22;29(2):405-418. [doi: 10.1007/s10115-010-0351-7]
79. Dao T, Marin F, Bensahel H, Ho Ba Tho MC. Computer-aided decision system for the clubfeet deformities. Adv Exp Med Biol 2011;696:623-635. [doi: 10.1007/978-1-4419-7046-6_64] [Medline: 21431604]
80. Cao F, Sun X, Wang X, Li B, Li J, Pan Y. Ontology-based knowledge management for personalized adverse drug events detection. Stud Health Technol Inform 2011;169:699-703. [Medline: 21893837]
81. Lee CS, Wang MH. A fuzzy expert system for diabetes decision support application. IEEE Trans Syst Man Cybern B Cybern 2011 Feb;41(1):139-153. [doi: 10.1109/TSMCB.2010.2048899] [Medline: 20501347]
82. Riaño D, Real F, López-Vallverdú JA, Campana F, Ercolani S, Mecocci P, et al. An ontology-based personalization of health-care knowledge to support clinical decisions for chronically ill patients. J Biomed Inform 2012 Jun;45(3):429-446 [FREE Full text] [doi: 10.1016/j.jbi.2011.12.008] [Medline: 22269224]
83. Bright TJ, Yoko Furuya E, Kuperman GJ, Cimino JJ, Bakken S. Development and evaluation of an ontology for guiding appropriate antibiotic prescribing. J Biomed Inform 2012 Feb;45(1):120-128 [FREE Full text] [doi: 10.1016/j.jbi.2011.10.001] [Medline: 22019377]
84. Koutkias V, Kilintzis V, Stalidis G, Lazou K, Niès J, Durand-Texte L, et al. Knowledge engineering for adverse drug event prevention: on the design and development of a uniform, contextualized and sustainable knowledge-based framework. J Biomed Inform 2012 Jun;45(3):495-506 [FREE Full text] [doi: 10.1016/j.jbi.2012.01.007] [Medline: 22326287]
85. Grando A, Farrish S, Boyd C, Boxwala A. Ontological approach for safe and effective polypharmacy prescription. AMIA Annu Symp Proc 2012;2012:291-300 [FREE Full text] [Medline: 23304299]
86. Chniti A, Boussadi A, Degoulet P, Albert P, Charlet J. Pharmaceutical validation of medication orders using an OWL Ontology and Business Rules. Stud Health Technol Inform 2012;180:1224-1226. [Medline: 22874408]
87. Paterno MD, Goldberg HS, Simonaitis L, Dixon BE, Wright A, Rocha BH, et al. Using a service oriented architecture approach to clinical decision support: performance results from two CDS Consortium demonstrations. AMIA Annu Symp Proc 2012;2012:690-698 [FREE Full text] [Medline: 23304342]
88. Yao W, Kumar A. CONFlexFlow: Integrating Flexible clinical pathways into clinical decision support systems using context and rules. Decision Support Systems 2013 May;55(2):499-515. [doi: 10.1016/j.dss.2012.10.008]
89. Artetxe A, Sanchez E, Toro C, Sanín C, Szczerbicki E, Graña M, et al. Impact of reflexive ontologies in semantic clinical decision support systems. Cybernetics and Systems 2013 Mar;44(2-3):187-203. [doi: 10.1080/01969722.2013.762256]

90. Farkash A, Timm JTE, Waks Z. A model-driven approach to clinical practice guidelines representation and evaluation using standards. Stud Health Technol Inform 2013;192:200-204. [Medline: 23920544]
91. Sáez C, Bresó A, Vicente J, Robles M, García-Gómez JM. An HL7-CDA wrapper for facilitating semantic interoperability to rule-based Clinical Decision Support Systems. Comput Methods Programs Biomed 2013 Mar;109(3):239-249. [doi: 10.1016/j.cmpb.2012.10.003] [Medline: 23199936]
92. Corrigan D, Taweel A, Fahey T, Arvanitis T, Delaney B. An ontological treatment of clinical prediction rules implementing the Alvarado score. Stud Health Technol Inform 2013;186:103-107. [Medline: 23542977]
93. Shojanoori R, Juric R. Semantic remote patient monitoring system. Telemed J E Health 2013 Feb;19(2):129-136. [doi: 10.1089/tmj.2012.0128] [Medline: 23363406]
94. Wilk S, Michalowski W, O'Sullivan D, Farion K, Sayyad-Shirabad J, Kuziemsky C, et al. A task-based support architecture for developing point-of-care clinical decision support systems for the emergency department. Methods Inf Med 2013;52(1):18-32 [FREE Full text] [doi: 10.3414/ME11-01-0099] [Medline: 23232759]
95. Yılmaz ?, Erdur RC, Türksever M. SAMS--a systems architecture for developing intelligent health information systems. J Med Syst 2013 Dec 7;37(6):9989. [doi: 10.1007/s10916-013-9989-5] [Medline: 24197356]
96. Bau C, Chen R, Huang C. Construction of a clinical decision support system for undergoing surgery based on domain ontology and rules reasoning. Telemed J E Health 2014 May;20(5):460-472 [FREE Full text] [doi: 10.1089/tmj.2013.0221] [Medline: 24730353]
97. Wang H, Zhou T, Tian L, Qian Y, Li J. Creating hospital-specific customized clinical pathways by applying semantic reasoning to clinical data. J Biomed Inform 2014 Dec;52:354-363 [FREE Full text] [doi: 10.1016/j.jbi.2014.07.017] [Medline: 25109270]
98. Sesen MB, Peake MD, Banares-Alcantara R, Tse D, Kadir T, Stanley R, et al. Lung Cancer Assistant: a hybrid clinical decision support application for lung cancer care. J R Soc Interface 2014 Sep 06;11(98):20140534 [FREE Full text] [doi: 10.1098/rsif.2014.0534] [Medline: 24990290]
99. Gallerani M, Pelizzola D, Pivanti M, Boni M, Lamma E, Bellodi E. Appropriateness of Repeated Execution of Laboratory Examinations: A CDSS Approach. 2014 Presented at: IEEE International Conference on Healthcare Informatics; September 15-17, 2014; Verona, Italy. [doi: 10.1109/ICHI.2014.29]
100. Stewart SA, Abidi S, Parker L, Bernstein M, Abidi SSR. Clinical guideline-driven personalized self-management diary for paediatric cancer survivors. Stud Health Technol Inform 2014;205:18-22. [Medline: 25160137]
101. Shen Y, Colloc J, Jacquet-Andrieu A, Lei K. Emerging medical informatics with case-based reasoning for aiding clinical decision in multi-agent system. J Biomed Inform 2015 Aug;56:307-317 [FREE Full text] [doi: 10.1016/j.jbi.2015.06.012] [Medline: 26133480]
102. Delaney BC, Curcin V, Andreasson A, Arvanitis TN, Bastiaens H, Corrigan D, et al. Translational Medicine and Patient Safety in Europe: TRANSFoRm--Architecture for the Learning Health System in Europe. Biomed Res Int 2015;2015:961526 [FREE Full text] [doi: 10.1155/2015/961526] [Medline: 26539547]
103. Jafarpour B, Abidi SR, Ahmad AM, Abidi SSR. INITIATE: An Intelligent Adaptive Alert Environment. Stud Health Technol Inform 2015;216:285-289. [Medline: 26262056]
104. Robles-Bykbaev V, López-Nores M, Pazos-Arias J, Quisi-Peralta D, García-Duque J. An Ecosystem of Intelligent ICT Tools for Speech-Language Therapy Based on a Formal Knowledge Model. Stud Health Technol Inform 2015;216:50-54. [Medline: 26262008]
105. Wilk S, Kezadri-Hamiaz M, Rosu D, Kuziemsky C, Michalowski W, Amyot D, et al. Using Semantic Components to Represent Dynamics of an Interdisciplinary Healthcare Team in a Multi-Agent Decision Support System. J Med Syst 2016 Feb 21;40(2):42. [doi: 10.1007/s10916-015-0375-3] [Medline: 26590980]
106. Abidi SR, Cox J, Abusharekh A, Hashemian N, Abidi SSR. A Digital Health System to Assist Family Physicians to Safely Prescribe NOAC Medications. Stud Health Technol Inform 2016;228:519-523. [Medline: 27577437]
107. Zhang Y, Tian Y, Zhou T, Araki K, Li J. Integrating HL7 RIM and ontology for unified knowledge and data representation in clinical decision support systems. Comput Methods Programs Biomed 2016 Jan;123:94-108. [doi: 10.1016/j.cmpb.2015.09.020] [Medline: 26474836]
108. Goldberg HS, Paterno MD, Grundmeier RW, Rocha BH, Hoffman JM, Tham E, et al. Use of a remote clinical decision support service for a multicenter trial to implement prediction rules for children with minor blunt head trauma. Int J Med Inform 2016 Mar;87:101-110. [doi: 10.1016/j.ijmedinf.2015.12.002] [Medline: 26806717]
109. Marco-Ruiz L, Pedrinaci C, Maldonado JA, Panziera L, Chen R, Bellika JG. Publication, discovery and interoperability of Clinical Decision Support Systems: A Linked Data approach. J Biomed Inform 2016 Aug;62:243-264 [FREE Full text] [doi: 10.1016/j.jbi.2016.07.011] [Medline: 27401856]
110. Zhang Y, Gou L, Zhou T, Lin D, Zheng J, Li Y, et al. An ontology-based approach to patient follow-up assessment for continuous and personalized chronic disease management. J Biomed Inform 2017 Aug;72:45-59 [FREE Full text] [doi: 10.1016/j.jbi.2017.06.021] [Medline: 28676255]
111. Shang Y, Wang Y, Gou L, Wu C, Zhou T, Li J. Development of a Service-Oriented Sharable Clinical Decision Support System Based on Ontology for Chronic Disease. Stud Health Technol Inform 2017;245:1153-1157. [Medline: 29295283]

112. Chen R, Jiang HQ, Huang C, Bau C. Clinical Decision Support System for Diabetes Based on Ontology Reasoning and TOPSIS Analysis. J Healthc Eng 2017;2017:4307508 [FREE Full text] [doi: 10.1155/2017/4307508] [Medline: 29312655]
113. Kopanitsa G. Integration of Hospital Information and Clinical Decision Support Systems to Enable the Reuse of Electronic Health Record Data. Methods Inf Med 2018 Jan 24;56(03):238-247. [doi: 10.3414/me16-01-0057]
114. Abidi S. A Knowledge-Modeling Approach to Integrate Multiple Clinical Practice Guidelines to Provide Evidence-Based Clinical Decision Support for Managing Comorbid Conditions. J Med Syst 2017 Oct 26;41(12):193. [doi: 10.1007/s10916-017-0841-1] [Medline: 29076113]
115. Shen Y, Yuan K, Chen D, Colloc J, Yang M, Li Y, et al. An ontology-driven clinical decision support system (IDDAP) for infectious disease diagnosis and antibiotic prescription. Artif Intell Med 2018 Mar;86:20-32. [doi: 10.1016/j.artmed.2018.01.003] [Medline: 29433958]
116. Nakawala H, Ferrigno G, De Momi E. Development of an intelligent surgical training system for Thoracentesis. Artif Intell Med 2018 Jan;84:50-63 [FREE Full text] [doi: 10.1016/j.artmed.2017.10.004] [Medline: 29169646]
117. Séroussi B, Guézennec G, Lamy J, Muro N, Larburu N, Sekar BD, et al. Reconciliation of multiple guidelines for decision support: a case study on the multidisciplinary management of breast cancer within the DESIREE project. AMIA Annu Symp Proc 2017;2017:1527-1536 [FREE Full text] [Medline: 29854222]
118. Winter A, Stäubert S, Ammon D, Aiche S, Beyan O, Bischoff V, et al. Smart Medical Information Technology for Healthcare (SMITH). Methods Inf Med 2018 Jul;57(S 01):e92-e105 [FREE Full text] [doi: 10.3414/ME18-02-0004] [Medline: 30016815]
119. Seroussi B, Lamy J, Muro N, Larburu N, Sekar BD, Guézennec G, et al. Implementing Guideline-Based, Experience-Based, and Case-Based Approaches to Enrich Decision Support for the Management of Breast Cancer Patients in the DESIREE Project. Stud Health Technol Inform 2018;255:190-194. [Medline: 30306934]
120. Jin W, Kim DH. Design and Implementation of e-Health System Based on Semantic Sensor Network Using IETF YANG. Sensors (Basel) 2018 Feb 20;18(2):629 [FREE Full text] [doi: 10.3390/s18020629] [Medline: 29461493]
121. Román-Villarán E, Pérez-Leon FP, Escobar-Rodriguez GA, Martínez-García A, Álvarez-Romero C, Parra-Calderón CL. An Ontology-Based Personalized Decision Support System for Use in the Complex Chronically Ill Patient. Stud Health Technol Inform 2019 Aug 21;264:758-762. [doi: 10.3233/SHTI190325] [Medline: 31438026]
122. Semenov I, Osenev R, Gerasimov S, Kopanitsa G, Denisov D, Andreychuk Y. Experience in Developing an FHIR Medical Data Management Platform to Provide Clinical Decision Support. Int J Environ Res Public Health 2019 Dec 20;17(1):73 [FREE Full text] [doi: 10.3390/ijerph17010073] [Medline: 31861851]
123. Jafarpour B, Raza Abidi S, Van Woensel W, Raza Abidi SS. Execution-time integration of clinical practice guidelines to provide decision support for comorbid conditions. Artif Intell Med 2019 Mar;94:117-137. [doi: 10.1016/j.artmed.2019.02.003] [Medline: 30871678]
124. El-Sappagh S, Ali F, Hendawi A, Jang J, Kwak K. A mobile health monitoring-and-treatment system based on integration of the SSN sensor ontology and the HL7 FHIR standard. BMC Med Inform Decis Mak 2019 May 10;19(1):97 [FREE Full text] [doi: 10.1186/s12911-019-0806-z] [Medline: 31077222]
125. Maldonado JA, Marcos M, Fernández-Breis JT, Giménez-Solano VM, Legaz-García MDC, Martínez-Salvador B. CLIN-IK-LINKS: A platform for the design and execution of clinical data transformation and reasoning workflows. Comput Methods Programs Biomed 2020 Dec;197:105616. [doi: 10.1016/j.cmpb.2020.105616] [Medline: 32629294]
126. Regier R, Gurjar R, Rocha RA. A clinical rule editor in an electronic medical record setting: development, design, and implementation. AMIA Annu Symp Proc 2009 Nov 14;2009:537-541 [FREE Full text] [Medline: 20351913]
127. Sordo M, Rocha BH, Morales AA, Maviglia SM, Oglio ED, Fairbanks A, et al. Modeling decision support rule interactions in a clinical setting. Stud Health Technol Inform 2013;192:908-912. [Medline: 23920690]
128. Bizer C, Heath T, Berners-Lee T. Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems 2009;5(3):1-22 [FREE Full text] [doi: 10.4018/jswis.2009081901]
129. Jing X, Min H, Gong Y. A clinical decision support system (CDSS) ontology to facilitate portable vaccination CDSS rules: preliminary results. 2021 Presented at: AMIA 2021 Annual Symposium; October 30-November 3, 2021; San Diego, CA.
130. Collen MF. Origins of medical informatics. West J Med 1986 Dec;145(6):778-785 [FREE Full text] [Medline: 3544507]
131. McCray AT. The Theoretical Basis of Medical Information Science. (Reflections on Marsden S. Blois' paper on the proper use of man and machines). Yearb Med Inform 1999(1):280-282. [Medline: 27699383]
132. Marco-Ruiz L, Budrionis A, Yigzaw KYY, Bellika JG. Interoperability Mechanisms of Clinical Decision Support Systems: A Systematic Review. In: Proceedings of the 14th Scandinavian Conference on Health Informatics. 2016 Presented at: 14th Scandinavian Conference on Health Informatics; April 6-7, 2016; Gothenburg, Sweden p. 13-21.
133. Marco-Ruiz L, Bellika JG. Semantic Interoperability in Clinical Decision Support Systems: A Systematic Review. Stud Health Technol Inform 2015;216:958. [Medline: 26262260]
134. Jess, the Rule Engine for the Java Platform. Sandia National Laboratories. URL: https://en.wikipedia.org/wiki/Jess_(programming_language) [accessed 2022-12-21]
135. Apache Jena. The Apache Software Foundation. URL: https://jena.apache.org/ [accessed 2022-12-21]
136. Drools. URL: https://www.drools.org/ [accessed 2022-12-21]

137. Parsia B, Matentzoglu N, Gonçalves RS, Glimm B, Steigmiller A. The OWL Reasoner Evaluation (ORE) 2015 Competition Report. J Autom Reason 2017;59(4):455-482 [FREE Full text] [doi: 10.1007/s10817-017-9406-8] [Medline: 30069067]
138. Sirin E, Parsia B, Grau B, Kalyanpur A, Katz Y. Pellet: A practical OWL-DL reasoner. Journal of Web Semantics 2007 Jun;5(2):51-53 [FREE Full text] [doi: 10.1016/j.websem.2007.03.004]
139. O'Connor M, Knublauch H, Tu S. Supporting Rule System Interoperability on the Semantic Web with SWRL. In: Gil Y, Motta E, Benjamins VR, Musen MA, editors. The Semantic Web – ISWC 2005. ISWC 2005. Lecture Notes in Computer Science, vol 3729. Berlin, Heidelberg: Springer; 2005:974-986.
140. Musen MA, Protégé Team. The Protégé Project: A Look Back and a Look Forward. AI Matters 2015 Jun;1(4):4-12 [FREE Full text] [doi: 10.1145/2757001.2757003] [Medline: 27239556]
141. Moreno-Conde A, Moner D, Cruz WDD, Santos MR, Maldonado JA, Robles M, et al. Clinical information modeling processes for semantic interoperability of electronic health records: systematic review and inductive analysis. J Am Med Inform Assoc 2015 Jul;22(4):925-934 [FREE Full text] [doi: 10.1093/jamia/ocv008] [Medline: 25796595]
142. Jing X, Kay S, Marley T, Hardiker NR, Cimino JJ. Incorporating personalized gene sequence variants, molecular genetics knowledge, and health knowledge into an EHR prototype based on the Continuity of Care Record standard. J Biomed Inform 2012 Feb;45(1):82-92 [FREE Full text] [doi: 10.1016/j.jbi.2011.09.001] [Medline: 21946299]
143. Jing X, Kay S, Marley T, Hardiker NR. Integration of an OWL-DL knowledge base with an EHR prototype and providing customized information. J Med Syst 2014 Sep;38(9):75. [doi: 10.1007/s10916-014-0075-4] [Medline: 24997857]

Abbreviations

ACM: Association for Computing Machinery
API: application programming interface
CDSS: clinical decision support system
EHR: electronic health record
EMR: electronic medical record
HIS: hospital information system
HL7: Health Level 7
NAHD: Nursing & Allied Health Database
OWL: Web Ontology Language
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RDF: resource description framework
SNOMED CT: Systematized Nomenclature of Medicine-Clinical Terms
SQWRL: Semantic Query-Enhanced Web Rule Language
SWRL: Semantic Web Rule Language

Edited by A Benis; submitted 28.09.22; peer-reviewed by M Bestek, J Schaaf; comments to author 22.10.22; revised version received 16.11.22; accepted 18.12.22; published 19.01.23.

Please cite as:
Jing X, Min H, Gong Y, Biondich P, Robinson D, Law T, Nohr C, Faxvaag A, Rennert L, Hubig N, Gimbel R
Ontologies Applied in Clinical Decision Support System Rules: Systematic Review
JMIR Med Inform 2023;11:e43053
URL: https://medinform.jmir.org/2023/1/e43053
doi:10.2196/43053
PMID:36534739

Review

The Current Status of Secondary Use of Claims, Electronic Medical Records, and Electronic Health Records in Epidemiology in Japan: Narrative Literature Review

Yang Zhao[1], PhD; Tadashi Tsubota[1], PhD

Audit & Assurance Deloitte Analytics R&D, Deloitte Touche Tohmatsu LLC, Tokyo, Japan

Corresponding Author:
Yang Zhao, PhD
Audit & Assurance Deloitte Analytics R&D
Deloitte Touche Tohmatsu LLC
Marunouchi Nijubashi Building
3-2-3 Marunouchi, Chiyoda-ku
Tokyo, 1008360
Japan
Phone: 81 80 9350 0848
Email: yang1.zhao@tohmatsu.co.jp

Abstract

Background: Real-world data, such as claims, electronic medical records (EMRs), and electronic health records (EHRs), are increasingly being used in clinical epidemiology. Understanding the current status of existing approaches can help in designing high-quality epidemiological studies.

Objective: We conducted a comprehensive narrative literature review to clarify the secondary use of claims, EMRs, and EHRs in clinical epidemiology in Japan.

Methods: We searched peer-reviewed publications in PubMed from January 1, 2006, to June 30, 2021 (the date of search), which met the following 3 inclusion criteria: involvement of claims, EMRs, EHRs, or medical receipt data; mention of Japan; and published from January 1, 2006, to June 30, 2021. Eligible articles that met any of the following 6 exclusion criteria were filtered: review articles; non–disease-related articles; articles in which the Japanese population is not the sample; articles without claims, EMRs, or EHRs; full text not available; and articles without statistical analysis. Investigations of the titles, abstracts, and full texts of eligible articles were conducted automatically or manually, from which 7 categories of key information were collected. The information included organization, study design, real-world data type, database, disease, outcome, and statistical method.

Results: A total of 620 eligible articles were identified for this narrative literature review. The results of the 7 categories suggested that most of the studies were conducted by academic institutes (n=429); the cohort study was the primary design that longitudinally measured outcomes of proper patients (n=533); 594 studies used claims data; the use of databases was concentrated in well-known commercial and public databases; infections (n=105), cardiovascular diseases (n=100), neoplasms (n=78), and nutritional and metabolic diseases (n=75) were the most studied diseases; most studies have focused on measuring treatment patterns (n=218), physiological or clinical characteristics (n=184), and mortality (n=137); and multivariate models were commonly used (n=414). Most (375/414, 90.6%) of these multivariate modeling studies were performed for confounder adjustment. Logistic regression was the first choice for assessing many of the outcomes, with the exception of hospitalization or hospital stay and resource use or costs, for both of which linear regression was commonly used.

Conclusions: This literature review provides a good understanding of the current status and trends in the use of claims, EMRs, and EHRs data in clinical epidemiology in Japan. The results demonstrated appropriate statistical methods regarding different outcomes, Japan-specific trends of disease areas, and the lack of use of artificial intelligence techniques in existing studies. In the future, a more precise comparison of relevant domestic research with worldwide research will be conducted to clarify the Japan-specific status and challenges.

(JMIR Med Inform 2023;11:e39876) doi:10.2196/39876

KEYWORDS

claims; electronic medical records; EMRs; electronic health records; EHRs; epidemiology; narrative literature review

Introduction

Background

Medical claims data, electronic medical records (EMRs), and electronic health records (EHRs) are familiar sources of real-world data (RWD). They are often used secondarily to complement limitations in clinical trials. For example, they can characterize patient subgroups that are excluded from clinical trials by following eligibility criteria such as comorbidities or age. Findings obtained through long-term, naturalistic observations of a large and diverse patient population can be easily generalized to other populations. Other advantages are that these data have high external validity, a single data source can be used for different study purposes, and prospective data collection is not required.

Claims data are electronic records of transactions between patients and health care providers. They include information on bills (claims) submitted by providers (hospitals, clinics, and pharmacies) to third-party payers (health insurance associations). There are already some large-scale commercial and nonprofit claims databases available in Japan [1-6] that aggregate information from multiple health care providers for secondary use**.** Recently, the EMR and EHR data have become widely available. The EMR data are the details of the encounters with patients recorded by physicians through EMR systems. They contain rich clinical information such as laboratory test results, diagnostic images, pathology findings, and patient symptoms. As different facilities may use different EMR systems, domestic EMR data are currently available from ≥1 medical institution. The EHR data are electronic records of all health-related information of individual patients created and managed by clinical professionals, which can be shared and used among various medical facilities. Current EHR databases in Japan include both patient claims data and medical records.

In recent years, claims data, EMRs, and EHRs have been increasingly used in clinical epidemiology studies. Such studies include cost-effectiveness analysis of drugs (including disease burden and assessment of medical technology), risk factor analysis, investigation of the actual status of drugs (including preclinical feasibility valuation, marketability study, and detection of prescription patterns), and evaluation of drug efficacy in actual clinical practice. Because these data are not designed for research purposes, the secondary use requires an understanding of their limitations and the ability to generate clinical questions, epidemiological skills to construct a study design, and statistical skills to analyze retrospective observational data. Previous approaches have addressed the limitations and challenges of using these data [7-12]. Understanding their application status based on these advanced guidelines is essential. However, investigations of existing epidemiological studies based on these data are lacking.

Objective

We conducted a comprehensive narrative literature review to clarify the secondary use of claims, EMRs, and EHRs in clinical epidemiology in Japan. We focused on 7 categories of key information, including organization, study design, RWD type, database, disease, outcome, and statistical method. We expect that this review would help in the design of high-quality epidemiological studies.

Methods

Overview

This is a comprehensive narrative literature review that investigated the secondary use of claims data, EMRs, and EHRs in epidemiology in Japan. Referring to PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [13] and procedures used in previous review studies [14-18], we conducted this review by searching for biomedical articles in PubMed.

Information Source

We searched peer-reviewed publications that satisfied the eligibility criteria for this narrative literature review in PubMed from January 1, 2006, to June 30, 2021 (the date of search).

Search Strategy

Keywords used to search PubMed consisted of “real world,” “database,” “claim,” “receipt,” “administrative,” “emr,” “ehr,” “japan,” “electronic medical record,” “electronic health record,” and Medical Subject Headings (MeSH) terms including, “Electronic Health Record,” “Administrative Claims, Healthcare,” “Insurance Claim Review/statistics and numerical data,” and “Japan/epidemiology.” We initially identified related articles by using various combinations of these keywords. The details of the search string are available in Multimedia Appendix 1.

Eligibility Criteria

On the basis of the search strategy, we identified articles whose titles and abstracts satisfied the following three inclusion criteria: (1) involvement of claims, EMRs, EHRs, or medical receipt data; (2) mention of Japan; and (3) published from January 1, 2006, to June 30, 2021. Eligible articles were then filtered out by satisfying any of the following six exclusion criteria: (1) review articles; (2) non–disease-related articles; (3) articles in which the Japanese population is not the sample; (4) articles without claims, EMRs, or EHRs; (5) unavailability of full-text articles; and (6) articles without statistical analysis.

Selection Process

The second author (TT) conducted the article search based on the search strategy. Both authors jointly reviewed all searched publications and performed 2 rounds of screening to identify target eligible articles. In the first round, we removed duplicates and articles that met any of the 6 exclusion criteria by screening the titles and abstracts. Review articles were automatically identified by a section classification model [19] trained on the PubMed 200k data set [20], which classified sentences in the abstracts into 5 sections (introduction, objective, method, result, and conclusion). On the basis of the hypothesis that review articles do not have sentences describing the results, we considered those without result sentences as review articles and removed them from the target articles. Artificially, we filtered out articles that met the exclusion criteria (2)-(5). In the second round of screening, the first author (YZ) reviewed the full text

of the remaining articles and removed those that did not include statistical analysis. The 2 authors double checked the results to ensure accuracy and finalized the eligible articles.

Data Collection

Overview

Investigations of the titles, abstracts, and full texts were conducted for eligible articles, from which 7 categories of key information were collected. The information included organization, study design, RWD type, database, disease, outcome, and statistical methods. Details regarding the classifications for each category are provided in Multimedia Appendix 2.

Automated Data Extraction

Four of these categories, including organization, study design, RWD type, and disease, were automatically extracted by keywords matching on the titles and abstracts. Two authors coded the data collection together.

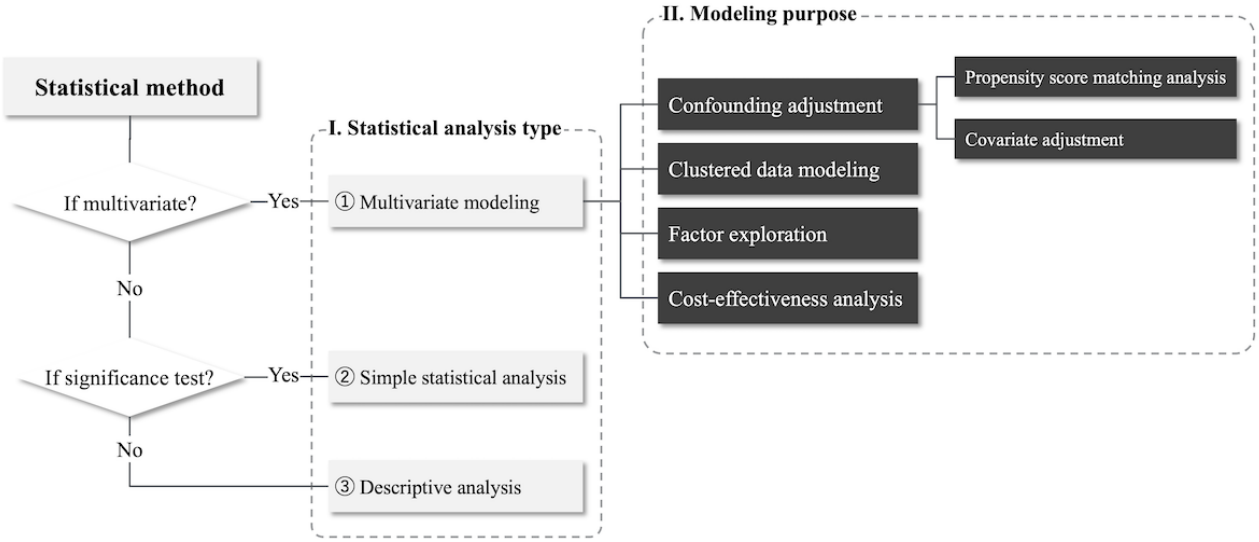
On the basis of authors' address information, organization was classified into 3 groups: "academic," "nonacademic," and "collaboration," which denote that a study was conducted by academia, enterprises (including pharmaceutical companies, biotechnology companies, medical device companies, voluntary associations, and other health care–related companies), or collaboration of academia and nonacademic enterprises, respectively. Study design information was extracted by matching sentences in the abstracts to the categories listed in Multimedia Appendix 2, which consists of cohort studies, case-control studies, case-crossover studies, and cross-sectional studies. Similarly, RWD-type information was extracted by matching sentences in the abstract with 3 keywords, including claims, EMRs, and EHRs. Disease information was classified according to tree codes C01-C26 of MeSH terms [21]. For articles without the corresponding MeSH terms, disease information was collected from their titles using MetaMap [22] and pyMeSHSim [23].

Manual Data Extraction

Subsequently, the first author (YZ) conducted a full-text investigation to collect information on the database, outcome, and statistical method used in the target articles. The second author (TT) cross-checked the results of this data collection.

Database information was collected directly from the full texts. For those articles that did not use a specific database, we categorized them uniformly according to their data source as "other database" or "municipal claims database," where "other database" indicates data from 1 or more medical facilities and "municipal claims database" indicates claims data provided by regional administrative agencies. Because there is no familiar way of categorizing outcomes for RWD studies, we defined 8 classifications of outcomes by referring to the article by Abaho et al [24]. The explanations for these classifications are detailed in Multimedia Appendix 2. We defined a hierarchical approach to collect information on statistical methods in the text. As shown in Figure 1, the method used in these articles was first categorized as multivariate modeling, simple statistical analysis, or descriptive analysis. Then, multivariate modeling was subdivided according to the purposes of confounding adjustment, clustered data modeling, factor exploration, or cost-effectiveness analysis, where confounding adjustment was further classified according to whether propensity score (PS) analysis was conducted.

Figure 1. A hierarchical approach for collecting information on the statistical method.



It should be noted that an article that focuses on multiple diseases, RWD types, study designs, databases, outcomes, or modeling purposes would be double counted for each classification to which it belongs.

Analysis

We performed a descriptive statistical analysis of the collected data by describing their counts and percentages. In addition, we calculated the percentages of outcomes and databases for each disease. The percentages of statistical methods used to assess different outcomes were also analyzed. All codes used for data

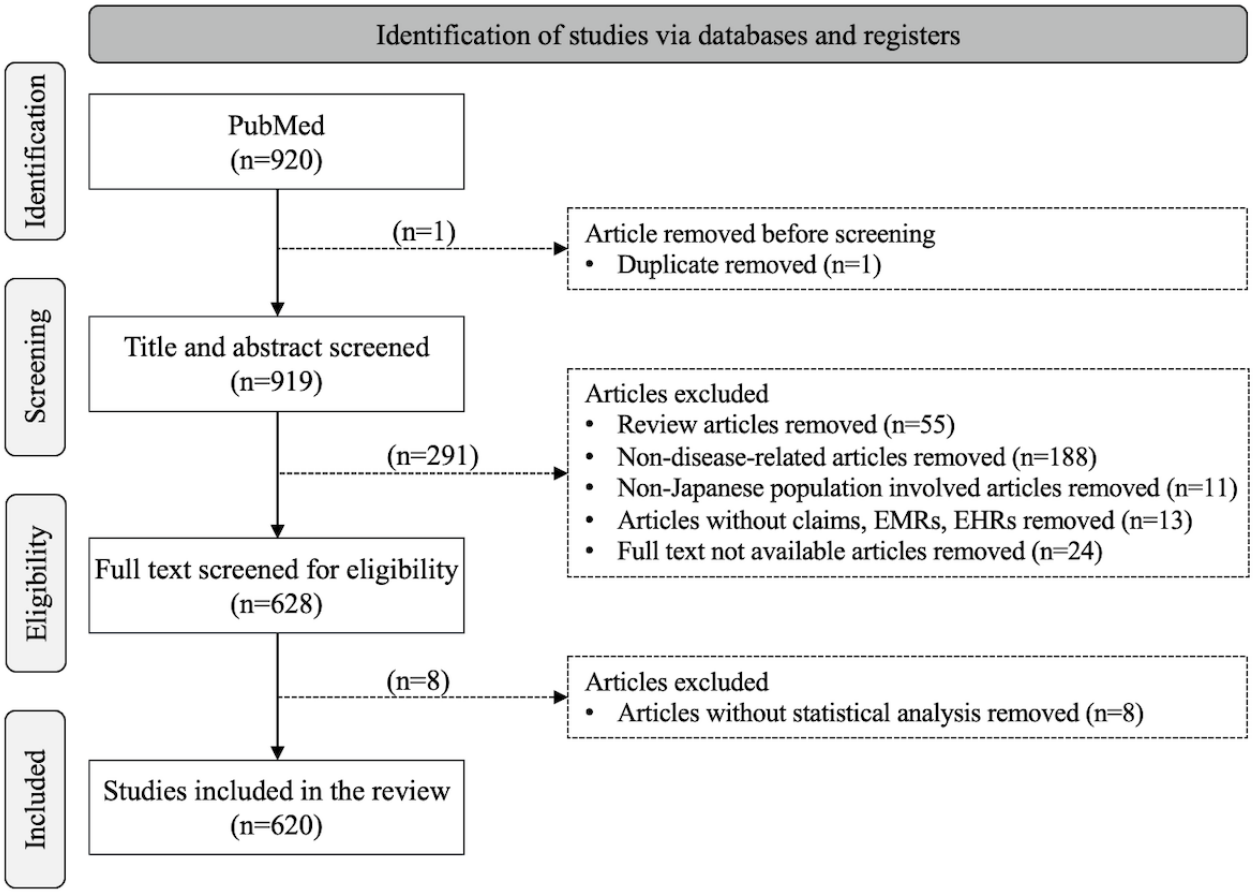
collection and descriptive analyses were performed using Python (version 3.8.8, 2021).

Results

Study Selection

A total of 620 eligible articles were identified for this narrative literature review. Figure 2 [13-18] illustrates the selection process and the results of each screening step. We also illustrate the publication years of these articles in Multimedia Appendix 3. The distribution indicated that 68.7% (426/620) of the articles were published after 2018, suggesting that the secondary use of the 3 RWD types in epidemiological research in Japan was prevalent in approximately the last 5 years.

Figure 2. Search and screening process [13-18]. EHR: electronic health record; EMR: electronic medical record.



Summary of Findings

Overview

We summarize the counts and percentages of information in the 7 categories and illustrate the top-ranked items for each category in Tables 1 and 2. All results are detailed in Multimedia Appendix 4. It should be noted that for an article with multiple diseases, data types, study designs, databases, outcomes, or modeling purposes, it was double counted in each classification to which it belongs. Therefore, the total percentage of these categories may not be 100%. The following subsections present the results for each category.

Table 1. Results of counts and percentages of the 7 categories (n=620).

Category	Count, n (%)
Organization	
Academic	429 (69.2)
Nonacademic	153 (24.7)
Collaboration	35 (5.6)
Study design	
Cohort study	533 (86)
Case-control study	30 (4.8)
Case-crossover study	23 (3.7)
Cross-sectional study	6 (1)
RWD[a] type	
Claim	594 (95.8)
EMR[b]	30 (4.8)
EHR[c]	4 (0.6)
Database	
JMDC[d]	181 (29.2)
DPC[e] database (MHLW[f])	141 (22.7)
MDV[g]	103 (16.6)
NDB[h]	65 (10.5)
Other databases	26 (4.2)
JROAD-DPC[i]	17 (2.7)
Municipal claims database	12 (1.9)
QIP[j]	10 (1.6)
Disease	
Infections	105 (16.9)
Cardiovascular diseases	100 (16.1)
Neoplasms	78 (12.6)
Nutritional and metabolic diseases	75 (12.1)
Digestive system diseases	68 (11)
Pathological conditions, signs and symptoms	63 (10.2)
Nervous system diseases	62 (10)
Musculoskeletal diseases	42 (6.8)
Mental disorders	38 (6.1)
Wounds and injuries	33 (5.3)
Male urogenital diseases	30 (4.8)
Respiratory tract diseases	27 (4.4)
Hemic and lymphatic diseases	16 (2.6)
Eye diseases	14 (2.3)
Skin and connective tissue diseases	10 (1.6)
Outcome	
Treatment patterns	218 (35.2)

Category	Count, n (%)
Physiological or clinical	184 (29.7)
Mortality	137 (22.1)
Resource use or costs	118 (19)
Hospitalization or hospital stay	107 (17.3)
Adverse events	97 (15.6)
Guideline adherence	32 (5.2)
Quality indicators	5 (0.8)
Statistical method	
Multivariate modeling	414 (66.8)
Simple statistical analysis	121 (19.5)
Descriptive analysis	85 (13.7)

[a]RWD: real-world data.
[b]EMR: electronic medical record.
[c]EHR: electronic health record.
[d]JMDC: Japan Medical Data Center Claims.
[e]DPC: diagnosis procedure combination.
[f]MHLW: Ministry of Health, Labour and Welfare.
[g]MDV: medical data vision.
[h]NDB: National Database of Health Insurance Claims and Specific Health Checkups of Japan.
[i]JROAD-DPC: Japanese Registry of All Cardiac and Vascular Disease-diagnosis procedure combination.
[j]QIP: Quality Indicator/Improvement Project.

Table 2. Results of modeling purposes as defined in Figure 1 and specific models used in the 414 multivariate modeling studies.

Category of multivariate modeling studies	Count (n=414), n (%)
Modeling purpose	
Confounding adjustment	375 (90.6)
Propensity score matching analysis	96 (23.2)
Covariate adjustment	279 (67.4)
Clustered data modeling	69 (16.7)
Factor exploration	68 (16.4)
Cost-effectiveness analysis	8 (1.9)
Specifical method	
Logistic regression	249 (60.1)
Cox proportional hazards regression	87 (21)
Linear regression	57 (13.8)
Poisson regression	23 (5.6)
GLM[a]	18 (4.3)

[a]GLM: generalized linear model.

Organization

In Table 1, the results of organization show that most (429/620, 69.2%) target articles were conducted by academics, whereas nonacademic firms preferred to collaborate with academic institutions (153/620, 24.7%).

Study Design

The results of study design show 86% (533/620) of the articles that performed cohort studies, whereas only a few (30/620, 4.8%) studies were case-control studies, cross-sectional studies (23/620, 3.7%), and case-crossover studies (6/620, 1%).

RWD Type

Most (594/620, 95.8%) studies used claims data. Only a small number (30/620, 4.8%) of studies used EMRs and (4/620, 0.6%) EHRs. According to the articles that used EMRs or EHRs, we found that these studies commonly collected EMRs or EHRs from private databases (1 or some specific hospitals), which did not have large patient populations.

Database

Table 1 shows the top-ranked databases (n≥10) used in the target articles. The Japan Medical Data Center Claims (JMDC) database, a well-known, large-scale commercial insurance-based claims database operated by JMDC Inc [3,4], was the most used database. JMDC was used in 29.2% (181/620) of the total articles. The second most used database is composed of claims data from diagnosis procedure combination (DPC) hospitals provided by the Ministry of Health, Labour and Welfare (MHLW) [25,26], which we called the DPC database (MHLW). A total of 22.7% (141/620) of articles used the DPC database (MHLW). Medical data vision (MDV) [5], another commercial hospital claims-based database, was used for 16.6% (103/620) of the total articles. Fourth in the ranking is the National Database of Health Insurance Claims and Specific Health Checkups of Japan (NDB) data, which was established by the MHLW in 2009, covering almost the whole population in Japan [1,2]. NDB was used in 10.5% (65/620) of the total articles.

Disease

According to the information on diseases in Table 1, we found that most studies have focused on infections (105/620, 16.9%), cardiovascular diseases (100/620, 16.1%), neoplasms (78/620, 12.6%), and nutritional and metabolic diseases (75/620, 12.1%). In addition, there were a number of studies on psychiatric disorders, indicated here as nervous system diseases (62/620, 10%) and mental disorders (38/620, 6.1%).

Outcome

The results of outcome show that treatment patterns (218/620, 35.2%), physiological or clinical outcomes (184/620, 29.7%), and mortality (137/620, 22.1%) were the most assessed outcomes. Comparatively, few (32/620, 5.2%) articles assessed guideline adherence. Only few studies measured quality indicators (5/620, 0.8%).

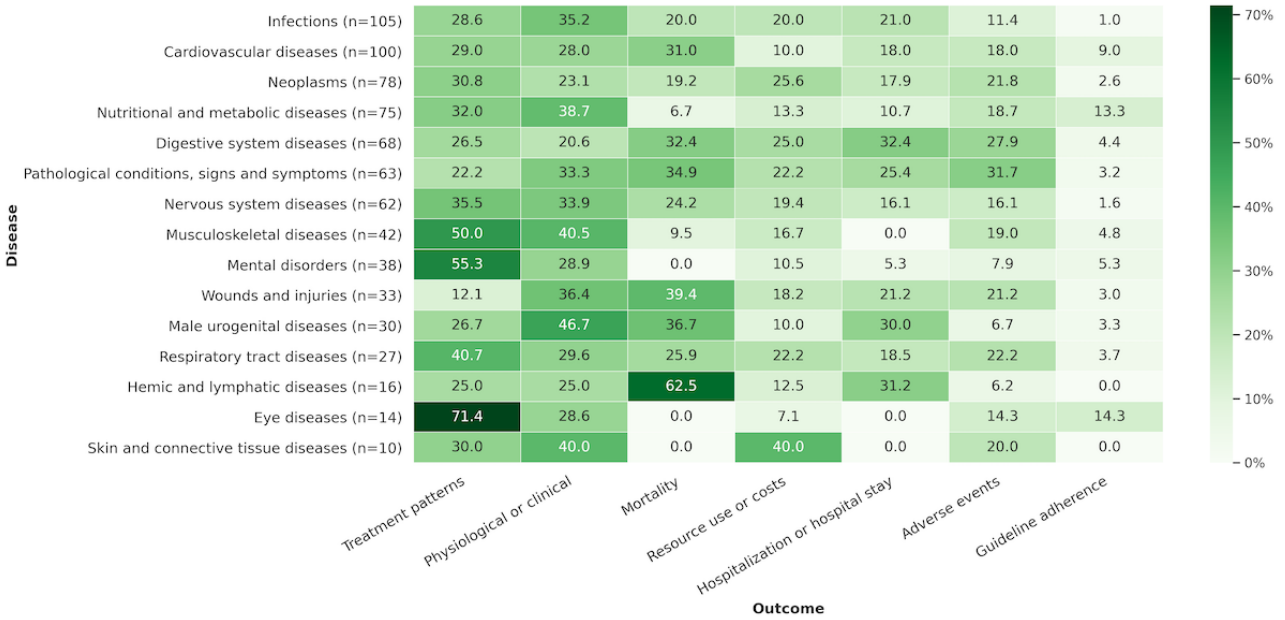
Statistical Method

Table 1 also suggests that most (414/620, 66.8%) studies were performed using multivariate modeling. In addition, we investigated the counts and percentages of modeling purposes (Figure 1) and specific models used in the 414 multivariate modeling studies in Table 2. The results show that most (375/414, 90.6%) of the multivariate modeling studies were performed for confounder adjustment. Some were conducted for clustered data modeling (69/414, 16.7%) and factor exploration (68/414, 16.4%). Two types of models were used for clustered data modeling: the generalized estimating equations (GEE) method and multilevel models. The GEE methods adjust for the clustering nature of the data and correctly estimate the SE of the estimated parameters. Multilevel models are often used with random effects to estimate the predictor effects for patients in specific clusters. Our results indicate a greater tendency to use multilevel regression (43/414, 10.4%) than GEE (26/414, 6.3%) in clustered data modeling studies. Only a few (8/414, 1.9%) studies analyzed cost-effectiveness. Regarding the specific models used in the multivariate modeling studies, logistic regression (249/414, 60.1%), Cox proportional hazards regression (87/414, 21%), and linear regression (57/414, 13.8%) were the most used.

Diseases and Outcomes

We investigated the percentage of each outcome measured for different diseases. As shown in Figure 3, most (10/14, 71%) studies on eye diseases have focused on assessing their treatment patterns. Similarly, a number of studies on mental disorders (21/38, 55%), musculoskeletal diseases (21/42, 50%), and respiratory tract diseases (11/27, 40%) have also focused on assessing treatment patterns. Among the studies on hemic and lymphatic diseases, mortality accounted for the highest percentage (10/16, 63%), whereas few studies assessed adverse events. Furthermore, mortality has not been assessed in studies of mental disorders, eye diseases, and skin and connective tissue diseases. In addition, no study has assessed hospitalization or hospital stay in musculoskeletal, eye, and skin and connective tissue diseases.

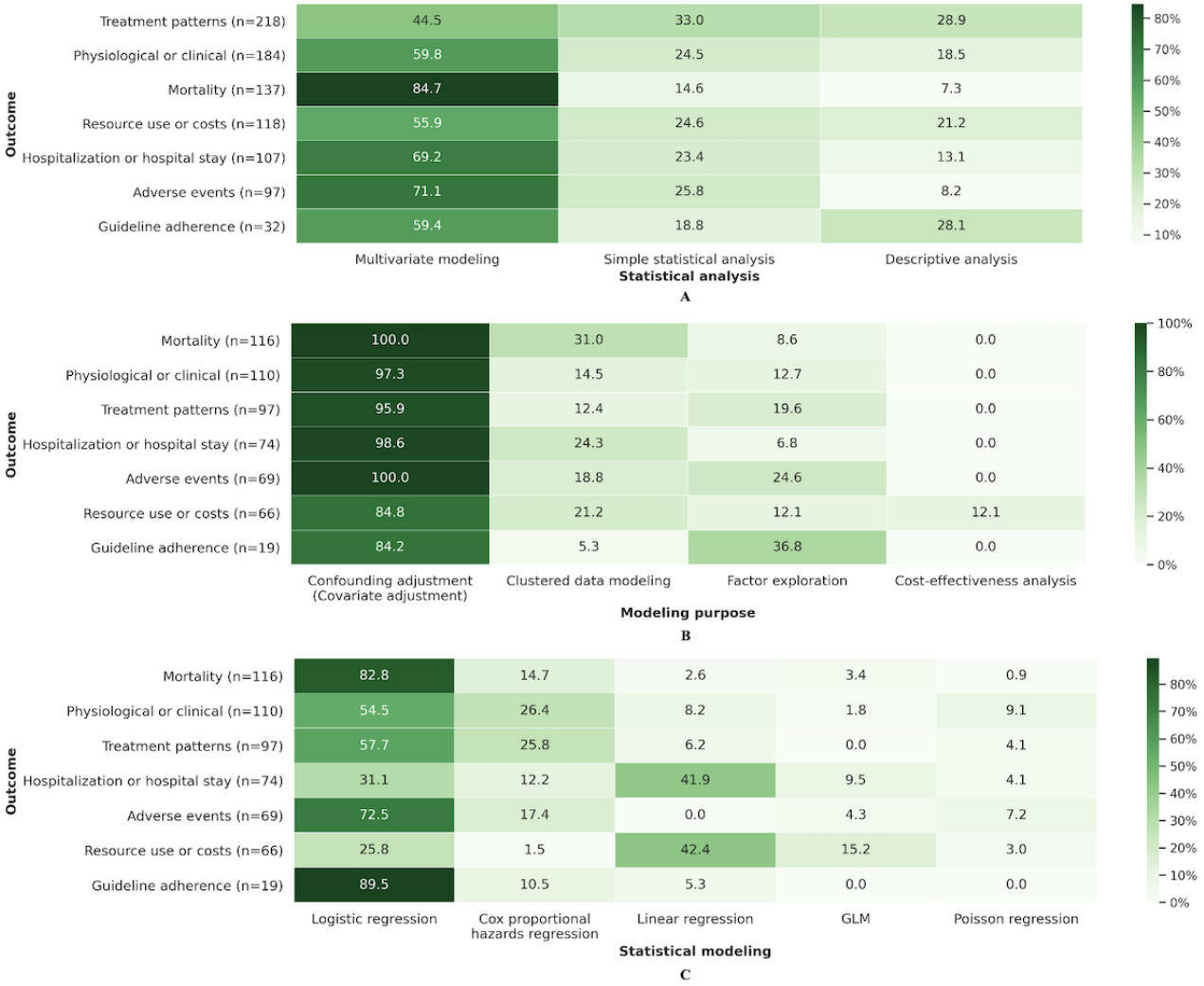
Figure 3. Percentages of outcomes in each disease.



Statistical Methods and Outcomes

We also calculated the percentages of statistical methods used to assess different outcomes. Figure 4A shows the percentages of the 3 types of statistical analyses used for each outcome; Figure 4B shows the percentages of multivariate modeling studies for different purposes for assessing these outcomes, and Figure 4C shows the percentage of each detailed multivariate model used for these outcomes. Multivariate modeling was used most frequently to assess mortality (116/137, 84.7%). Although the treatment patterns were the most assessed by the target studies (n=218), not many of them used multivariate modeling (97/218, 44.5%). Figure 4B indicates that almost all outcomes were measured with confounding adjustments. As shown in Figure 4C, logistic regression was the first choice for assessing mortality (96/116, 82.8%), physiological or clinical outcomes (60/110, 54.5%), treatment patterns (56/97, 58%), and guideline adherence (17/19, 90%). The results also suggest the use of Cox proportional hazards regression to assess these outcomes. In contrast, linear regression was the most commonly used model for assessing hospitalization or hospital stay (31/74, 42%) and resource use or costs (28/66, 42%).

Figure 4. (A) Percentages of statistical analysis types for each outcome, (B) modeling purposes for each outcome, and (C) specific models for each outcome. GLM: generalized linear model.



Discussion

Principal Findings

A comprehensive narrative literature review was conducted to understand the secondary use of nationwide claims data, EMRs data, and EHRs data in clinical epidemiology in Japan. On the basis of the search strategy and eligibility criteria, a total of 620 eligible articles were identified from PubMed between January 1, 2006, and June 30, 2021 (the date of search).

We quantified 7 categories of key information from these 620 eligible articles. The main findings were that (1) most of the research has been done by academic institutions, whereas nonacademic institutions tend to collaborate with academic institutions; (2) the cohort study was the major design that longitudinally measured outcomes of proper patients; (3) most studies used claims data; (4) the JMDC, DPC database (MHLW), MDV, and NDB were mostly used, whereas only a few studies used EMRs or EHRs from a single hospital or multiple hospitals, which do not have a large patient population; (5) the top rank of diseases studied in the current research were infections, cardiovascular diseases, neoplasms, and nutritional and metabolic diseases; (6) treatment patterns, physiological or clinical outcomes, and mortality were the most assessed in these articles; and (7) multivariate models were commonly used, during which logistic regression and linear regression were shown to be the first choice for analyzing categorical variables and continuous variables, respectively.

The findings on the percentage of outcomes for different diseases hint at the tendency of existing studies to examine different diseases. For some common, chronic, and psychiatric diseases, current studies tended to assess their treatment patterns, whereas for some sudden onset severe diseases, patient mortality and hospitalization or hospital stay were assessed more often. Existing studies have focused more on assessing treatment modalities, physiological or clinical outcomes, and mortality when targeting diseases such as infections, cardiovascular diseases, and neoplasms. Furthermore, although strong trends were detected between eye diseases and treatment patterns, hemic and lymphatic diseases versus mortality, and mental disorders versus mortality (Figure 3), it was difficult to draw any conclusions that reflect clinical importance because of the small sample size. However, these results indicated different distributions of outcomes measured in different diseases, from which we can learn the focus and shortcomings of the existing studies. In addition, the total number of studies measuring guideline adherence was relatively small (n=32). During this

period, 63% (20/32) of the studies were conducted on "cardiovascular diseases" and "nutritional and metabolic diseases." These results also revealed a relative lack of studies measuring guideline adherence in infections. We expect that RWD research on guideline adherence would receive more attention in future.

The percentage of databases used for different diseases implied the selection of databases for observing different diseases. The JMDC databases and DPC database (MHLW) showed opposite use trends in diseases, especially nutritional and metabolic diseases, musculoskeletal diseases, mental disorders, hemic and lymphatic diseases, eye diseases, and skin diseases.

According to the investigation of statistical methods used to assess different outcomes, multivariate models were the most commonly used in assessing mortality. Regardless of the outcome, multivariate modeling was accompanied by adjustments for various confounders (Figure 4B). Mortality, hospitalization or hospital stay, and resource use or costs have been analyzed using multilevel models or marginal models (eg, GEE) more than others. This implies that hospital-related outcomes tended to be assessed by models that took clustering into account. Logistic regression was the first choice for measuring many of the outcomes, with the exception of hospitalization or hospital stay and resource use or costs, for which linear regression was commonly used. Cox proportional hazards regression was suggested as the second choice when assessing mortality, physiological or clinical outcomes, and treatment patterns. Although the PS technique has been proven effective in balancing confounders between groups, it has not been widely used in existing studies. There is a relative preference for this technique in studies assessing mortality.

Comparison With Prior Work

In this subsection, we compare this review with 2 similar studies [27,28]. Hirose et al [27] conducted a narrative review of 68 studies on the secondary use of claims data in a specific database, NDB, from October 2016 to June 2019. They summarized 5 key pieces of information, including study design, research area, setting or sample, outcomes, and strengths and limitations. Subsequently, Fujinaga and Fukuoka [28] conducted a similar narrative review of 643 studies on the secondary use of claims data in 4 large-scale domestic databases: NDB, DPC database (MHLW), JMDC, and MDV, from January 2015 to October 2020, from which 3 categories of research type, design, and area were analyzed descriptively. Both studies used a classification of the journals in which the target articles were published to extract information about the research area [29]. These classifications mixed disciplinary categories, such as clinical medicine, pharmacology and pharmacy, pharmacology and toxicology, and immunology; disease categories, such as infectious diseases; and general categories, such as social sciences and public environmental health. In addition, only the primary outcomes were analyzed in these 2 studies. As a result, the distribution of articles in each category was summarized in these studies.

Because of the partial overlap in search periods, as well as the fact that PubMed was used for the search, there were some articles that were reviewed in both this study and these 2 prior studies. In contrast to these 2 studies, which used 1 or more specific claim databases without specifying a research area, our review investigated domestic epidemiological studies based on the secondary use of 3 types of RWD: claims, EMRs, and EHRs. A further difference is that we defined 7 categories for data collection to assess the status and trends of the existing studies. One of the novelties is that we classified the outcomes with reference to the paper by Abaho et al [24] paper and collected information on all the outcomes measured in the target articles. The advantage of this classification is that these outcomes are also applicable to clinical trial studies and can be automatically identified from biomedical articles [24]. Another innovative point is that we proposed a hierarchical approach to classify the statistical methods that appear in the target articles. For the results of the data collection, we summarized the distribution of the target articles in each category. Additional comparative analyses were performed for diseases versus outcomes (Figure 3), outcomes versus statistical methods (Figure 4), and diseases versus databases (Multimedia Appendix 5), which revealed trends in the assessment of outcomes across different diseases, trends of statistical methods used for different outcomes, and trends in database selection when analyzing different diseases. Moreover, our findings shed light on the focus and shortcomings of previous studies.

In addition, we identified several other review studies on the secondary use of RWD data [30-32]. The paper by Ferver et al [30] provided a narrative review of 1956 claims-based studies in 5 health care journals from 2000 to 2005 by summarizing the research types and areas. The paper by Hutchings et al [31] provided a systematic literature review of 18 studies to investigate the attitudes of relevant practitioners toward the secondary use and sharing of health administrative and clinical trial data. Schlegel et al [32] conducted a literature review of 941 studies on the secondary use of health care data in 2016 to select the best performing articles. We summarized these additional studies to understand other investigations on the secondary use of RWD data. Comparisons were not made because of the survey years or different research purposes.

Limitations

The first limitation of this review is that we only searched the literature in PubMed, which may have led to significant publication bias. Second, we only investigated studies conducted in Japan. In the future, a comparison of studies from other countries, such as the United States, will be necessary to understand the Japan-specific trends of such studies. In addition, searches of multiple electronic databases should be considered to reduce potential publication bias.

Future Directions

In this subsection, we discuss the future perspectives for the use of claims, EMRs, and EHRs in epidemiology in the Japanese context, in terms of the findings of this large narrative literature review.

Organization

Regarding collaborative aspects, with strong national promotion for RWD use and high level of interest from health care firms, collaborative research, involving multiple stakeholders and

academic researchers, is seen to be necessary to leverage academic results and accelerate clinical applications.

RWD Type

Notably, only a few studies have used EHRs. EHRs have not been widespread in Japan because of the high cost of implementation and the difficulties in bridging different EHR service vendors. With the promotion of "cloud-based EHR" development by the Japanese Ministry of Internal Affairs and Communications, EHRs are expected to become widely used in the future.

Disease

With regard to the disease trend detected in this review, we made a rough comparison with worldwide trends. As we did not find a quantitative survey of RWD research on different diseases, the worldwide trend was roughly estimated by counting the number of related publications for different diseases. We focused on the top-ranked disease areas identified in this review, including infections, cardiovascular diseases, and neoplasms. The number of publications for these diseases was obtained by searching for electronic databases, such as PubMed or PubMed Central with search keywords: combinations of "claims," "EHR," "EMR," to "infection," "cardiovascular disease," and "cancer." We retrieved 18,847 publications on cancer, 7517 publications on infections, and 6624 publications on cardiovascular diseases from PubMed. The same trend was detected in PubMed Central. According to these counts, we estimated that the worldwide trend of the disease examined in existing studies was cancer. In contrast, our results revealed a Japan-specific trend in the studies on infections.

It is important to note that the above counts may be subject to bias because we have not designed any eligibility criteria for the precise search of related publications worldwide. In the future, it will be necessary to compare relevant studies with those of other countries to clarify the Japan-specific status and challenges.

Statistical Method

On the basis of the statistical skills used in the eligible articles, we summarized the appropriate statistical methods for use under different conditions. First, to design simple statistical analyses, our findings suggest using Fisher's exact tests or chi-square test to compare categorical variables, and 2-tailed *t* test, ANOVA, and Mann-Whitney *U* test were used to compare continuous variables [33-36]. To evaluate variable change trends, the Cochran-Armitage test was used for categorical variables, whereas the Jonckheere-Terpstra test was used for continuous variables [37].

Suggestions for statistical methods to measure different outcomes are summarized in Table 3. For confounding adjustment, there are 2 methods: covariate adjustment and PS analysis. PS analysis is known to be an effective technique for balancing the patient backgrounds between the 2 groups across all putative risk factors or confounders [38-40]. However, referring to the study by Elze et al [41] that PS analysis is not necessarily superior to conventional covariate adjustment, we suggest selecting PS analysis with caution for confounder adjustment. Our findings also demonstrated that most existing studies used covariate adjustment (n=279) rather than PS analysis (n=96; Multimedia Appendix 4). In addition, hospital-based medical data are frequently clustered within medical centers or physicians. For instance, patients treated in a particular hospital may be more alike than those treated in another hospital because of differences in treatment policies. To model such clustered data, multilevel models with random effects have been suggested for use in estimating predictor effects for patients in specific clusters [42,43].

Table 3. Suggestions of statistical methods for measuring different outcomes.

Outcome	Method recommendation
Treatment patterns	Logistic regression, Cox proportional hazards regression
Physiological or clinical	Logistic regression, Cox proportional hazards regression
Mortality	Kaplan-Meier analysis, log-rank test, logistic regression, Cox proportional hazards regression
Hospitalization or hospital stay	Linear regression, GLM[a]
Adverse events	Logistic regression, Cox proportional hazards regression
Resource use or costs	Linear regression, GLM
Guideline adherence	Logistic regression
Quality indicators	Logistic regression

[a]GLM: generalized linear model.

In contrast, there were few studies on predictive machine learning models in this review (n=3; Multimedia Appendix 4). However, we roughly retrieved 2223 publications worldwide on PubMed by searching for the keywords of "claims," "EHR," "EMR," and "machine learning." Notably, we did not design any eligibility criteria for this study. The large difference in the number of articles indicates that epidemiological research based on claims, EMRs, and EHRs in Japan is backward in the use of artificial intelligence techniques.

Conclusions

This literature review provides a good understanding of the current status and trends in the use of claims, EMRs, and EHRs in clinical epidemiology in Japan. The results demonstrated appropriate statistical methods regarding different outcomes, Japan-specific trend of disease areas, and lack of use of artificial

intelligence techniques in existing studies. We hope that the results of this narrative review will provide useful information for researchers to design relevant studies. In the future, a more precise comparison of relevant domestic research with worldwide research will be conducted to clarify the Japan-specific status and challenges.

Data Availability

All data generated or analyzed during this study are included in this published paper and its multimedia appendices.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PubMed search string.
[DOCX File , 28 KB - medinform_v11i1e39876_app1.docx]

Multimedia Appendix 2
Explanation of the 7 categories of information.
[DOCX File , 32 KB - medinform_v11i1e39876_app2.docx]

Multimedia Appendix 3
Distribution of publication years of the 620 eligible articles.
[PNG File , 153 KB - medinform_v11i1e39876_app3.png]

Multimedia Appendix 4
Full results of counts and percentages of the 7 categories.
[XLSX File (Microsoft Excel File), 21 KB - medinform_v11i1e39876_app4.xlsx]

Multimedia Appendix 5
Percentages of databases used in each disease.
[PNG File , 602 KB - medinform_v11i1e39876_app5.png]

References

1. [NDB] Website for provision of information such as anonymous receipt information and anonymous specific health checkup information. Ministry of Health, Labor and Welfare. URL: https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/kenkou_iryou/iryouhoken/reseputo/index.html [accessed 2022-09-30]
2. Kubo S, Noda T, Myojin T, Nishioka Y, Higashino T, Matsui H, et al. National Database of Health Insurance Claims and Specific Health Checkups of Japan (NDB): outline and patient-matching technique. bioRxiv 2018 Apr 2. [doi: 10.1101/280008]
3. JMDC Claims Database. JMDC Inc. URL: https://www.jmdc.co.jp/en/jmdc-claims-database/ [accessed 2022-09-30]
4. Nagai K, Tanaka T, Kodaira N, Kimura S, Takahashi Y, Nakayama T. Data resource profile: JMDC claims databases sourced from Medical Institutions. J Gen Fam Med 2020 Nov;21(6):211-218 [FREE Full text] [doi: 10.1002/jgf2.367] [Medline: 33304714]
5. What makes MDV Database special? MDV. URL: https://en.mdv.co.jp/about-mdv-database/ [accessed 2022-09-30]
6. Laurent T, Simeone J, Kuwatsuru R, Hirano T, Graham S, Wakabayashi R, et al. Context and considerations for use of two Japanese real-world databases in japan: medical data vision and Japanese medical data center. Drugs Real World Outcomes 2022 Jun;9(2):175-187 [FREE Full text] [doi: 10.1007/s40801-022-00296-5] [Medline: 35304702]
7. Hiramatsu K, Barrett A, Miyata Y, PhRMA Japan Medical Affairs Committee Working Group 1. Current status, challenges, and future perspectives of real-world data and real-world evidence in Japan. Drugs Real World Outcomes 2021 Dec;8(4):459-480 [FREE Full text] [doi: 10.1007/s40801-021-00266-3] [Medline: 34148219]
8. Ishikawa KB. Medical big data for research use: current status and related issues. Japan Med Assoc J 2016 Sep;59(2-3):110-124 [FREE Full text] [Medline: 28299245]
9. Wakabayashi Y, Eitoku M, Suganuma N. Characterization and selection of Japanese electronic health record databases used as data sources for non-interventional observational studies. BMC Med Inform Decis Mak 2021 May 22;21(1):167 [FREE Full text] [doi: 10.1186/s12911-021-01526-6] [Medline: 34022876]
10. Kumamaru H, Fukuma S, Matsui H, Kawasaki R, Tokumasu H, Takahashi A, et al. Principles for the use of large-scale medical databases to generate real-world evidence. Annals Clin Epidemiol 2020;2(1):27-32. [doi: 10.37737/ace.2.1_27]

11. Hanada K, Akazawa M. Current status of real world data using for conducting cost-effectiveness assessments in Japan. Value Health 2018 Sep;21:S97. [doi: 10.1016/j.jval.2018.07.736]
12. Maeda H. The current status and future direction of clinical research in Japan from a regulatory perspective. Front Med (Lausanne) 2021;8:816921 [FREE Full text] [doi: 10.3389/fmed.2021.816921] [Medline: 35096908]
13. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
14. Watkins I, Xie B. eHealth literacy interventions for older adults: a systematic review of the literature. J Med Internet Res 2014 Nov 10;16(11):e225 [FREE Full text] [doi: 10.2196/jmir.3318] [Medline: 25386719]
15. Powers EM, Shiffman RN, Melnick ER, Hickner A, Sharifi M. Efficacy and unintended consequences of hard-stop alerts in electronic health record systems: a systematic review. J Am Med Inform Assoc 2018 Nov 01;25(11):1556-1566 [FREE Full text] [doi: 10.1093/jamia/ocy112] [Medline: 30239810]
16. Choudhury A, Asan O. Role of artificial intelligence in patient safety outcomes: systematic literature review. JMIR Med Inform 2020 Jul 24;8(7):e18599 [FREE Full text] [doi: 10.2196/18599] [Medline: 32706688]
17. Xie B, Tao C, Li J, Hilsabeck RC, Aguirre A. Artificial intelligence for caregivers of persons with Alzheimer's disease and related dementias: systematic literature review. JMIR Med Inform 2020 Aug 20;8(8):e18189 [FREE Full text] [doi: 10.2196/18189] [Medline: 32663146]
18. Duffy A, Christie GJ, Moreno S. The challenges toward real-world implementation of digital health design approaches: narrative review. JMIR Hum Factors 2022 Sep 09;9(3):e35693 [FREE Full text] [doi: 10.2196/35693] [Medline: 36083628]
19. Beltagy I, Lo K, Cohan A. SciBERT: a pretrained language model for scientific text. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019 Presented at: 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); Nov, 2019; Hong Kong, China. [doi: 10.18653/v1/D19-1371]
20. Dernoncourt F, Lee J. PubMed 200k RCT: a dataset for sequential sentence classification in medical abstracts. arXiv 2017. [doi: 10.48550/arXiv.1710.06071]
21. Welcome to Medical Subject Headings. National Library of Medicine. URL: https://www.nlm.nih.gov/mesh/meshhome.html [accessed 2022-09-30]
22. MetaMap - a tool for recognizing UMLS concepts in text. National Library of Medicine. URL: https://lhncbc.nlm.nih.gov/ii/tools/MetaMap.html [accessed 2022-09-30]
23. Luo Z, Shi M, Yang Z, Zhang H, Chen Z. pyMeSHSim: an integrative python package for biomedical named entity recognition, normalization, and comparison of MeSH terms. BMC Bioinformatics 2020 Jun 18;21(1):252 [FREE Full text] [doi: 10.1186/s12859-020-03583-6] [Medline: 32552728]
24. Abaho M, Bollegala D, Williamson P, Dodd S. Assessment of contextualised representations in detecting outcome phrases in clinical trials. Eur J Biomedl Inform 2022 Mar 13 [FREE Full text] [doi: 10.24105/EJBI.2021.17.9.53-65]
25. Yasunaga H, Matsui H, Horiguchi H, Fushimi K, Matsuda S. Clinical epidemiology and health services research using the diagnosis procedure combination database in Japan. Asian Pacific J Disease Manag 2015;7(1-2):19-24. [doi: 10.7223/apjdm.7.19]
26. Fushimi K, Hashimoto H, Imanaka Y, Kuwabara K, Horiguchi H, Ishikawa KB, et al. Functional mapping of hospitals by diagnosis-dominant case-mix analysis. BMC Health Serv Res 2007 Apr 10;7:50 [FREE Full text] [doi: 10.1186/1472-6963-7-50] [Medline: 17425788]
27. Hirose N, Ishimaru M, Morita K, Yasunaga H. A review of studies using the Japanese national database of health insurance claims and specific health checkups. Annals Clin Epidemiol 2020;2(1):13-26 [FREE Full text] [doi: 10.37737/ace.2.1_13]
28. Fujinaga J, Fukuoka T. A review of research studies using data from the administrative claims databases in Japan. Drugs Real World Outcomes 2022 Dec;9(4):543-550 [FREE Full text] [doi: 10.1007/s40801-022-00331-5] [Medline: 36107390]
29. Larsen PO, von Ins M. The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. Scientometrics 2010 Sep;84(3):575-603 [FREE Full text] [doi: 10.1007/s11192-010-0202-z] [Medline: 20700371]
30. Ferver K, Burton B, Jesilow P. The use of claims data in healthcare research. Open Public Health J 2009 Apr 02;2(1):11-24. [doi: 10.2174/1874944500902010011]
31. Hutchings E, Loomes M, Butow P, Boyle FM. A systematic literature review of researchers' and healthcare professionals' attitudes towards the secondary use and sharing of health administrative and clinical trial data. Syst Rev 2020 Oct 12;9(1):240 [FREE Full text] [doi: 10.1186/s13643-020-01485-5] [Medline: 33046097]
32. Schlegel DR, Ficheur G. Secondary use of patient data: review of the literature published in 2016. Yearb Med Inform 2017 Aug;26(1):68-71 [FREE Full text] [doi: 10.15265/IY-2017-032] [Medline: 29063536]
33. Sato Y, Miyashita M, Sato K, Fujimori K, Ishikawa KB, Horiguchi H, et al. End-of-life care for cancer patients in Japanese acute care hospitals: a nationwide retrospective administrative database survey. Jpn J Clin Oncol 2018 Oct 01;48(10):877-883. [doi: 10.1093/jjco/hyy117] [Medline: 30107588]

34. Nagano H, Takada D, Shin J, Morishita T, Kunisawa S, Imanaka Y. Hospitalization of mild cases of community-acquired pneumonia decreased more than severe cases during the COVID-19 pandemic. Int J Infect Dis 2021 May;106:323-328 [FREE Full text] [doi: 10.1016/j.ijid.2021.03.074] [Medline: 33794382]
35. Shirai T, Imanaka Y, Sekimoto M, Ishizaki T, QIP Ovarian Cancer Expert Group. Primary chemotherapy patterns for ovarian cancer treatment in Japan. J Obstet Gynaecol Res 2009 Oct;35(5):926-934. [doi: 10.1111/j.1447-0756.2009.01033.x] [Medline: 20149043]
36. Nakashima M, Takeuchi M, Kawakami K. Clinical outcomes of acute appendicitis during pregnancy: conservative management and appendectomy. World J Surg 2021 Jun;45(6):1717-1724. [doi: 10.1007/s00268-021-06010-w] [Medline: 33635341]
37. Yamashita Y, Morimoto T, Yoshikawa Y, Yaku H, Sumita Y, Nakai M, et al. Temporal trends in the practice pattern for venous thromboembolism in Japan: insight from JROAD-DPC. J Am Heart Assoc 2020 Jan 21;9(2):e014582 [FREE Full text] [doi: 10.1161/JAHA.119.014582] [Medline: 31918600]
38. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983;70(1):41-55 [FREE Full text] [doi: 10.1093/biomet/70.1.41]
39. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behav Res 2011 May;46(3):399-424 [FREE Full text] [doi: 10.1080/00273171.2011.568786] [Medline: 21818162]
40. Austin PC, Xin Yu AY, Vyas MV, Kapral MK. Applying propensity score methods in clinical research in neurology. Neurology 2021 Nov 02;97(18):856-863 [FREE Full text] [doi: 10.1212/WNL.0000000000012777] [Medline: 34504033]
41. Elze MC, Gregson J, Baber U, Williamson E, Sartori S, Mehran R, et al. Comparison of propensity score methods and covariate adjustment: evaluation in 4 cardiovascular studies. J Am Coll Cardiol 2017 Jan 24;69(3):345-357 [FREE Full text] [doi: 10.1016/j.jacc.2016.10.060] [Medline: 28104076]
42. Fukuda H, Sato D, Kato Y, Tsuruta W, Katsumata M, Hosoo H, et al. Comparing retreatments and expenditures in flow diversion versus coiling for unruptured intracranial aneurysm treatment: a retrospective cohort study using a real-world national database. Neurosurgery 2020 Jul 01;87(1):63-70. [doi: 10.1093/neuros/nyz377] [Medline: 31541237]
43. Miki R, Takeuchi M, Imai T, Seki T, Tanaka S, Nakamura M, et al. Association of intensive care unit admission and mortality in patients with acute myocardial infarction. J Cardiol 2019 Aug;74(2):109-115 [FREE Full text] [doi: 10.1016/j.jjcc.2019.01.007] [Medline: 30773390]

Abbreviations

DPC: diagnosis procedure combination
EHR: electronic health record
EMR: electronic medical record
GEE: generalized estimating equations
GLM: generalized linear model
JMDC: Japan Medical Data Center Claims
MDV: medical data vision
MeSH: Medical Subject Headings
MHLW: Ministry of Health, Labour and Welfare
NDB: National Database of Health Insurance Claims and Specific Health Checkups of Japan
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PS: propensity score
RWD: real-world data

Edited by J Hefner; submitted 26.05.22; peer-reviewed by CMJ Wong, J Shull; comments to author 10.08.22; revised version received 01.11.22; accepted 05.01.23; published 14.02.23.

Please cite as:
Zhao Y, Tsubota T
The Current Status of Secondary Use of Claims, Electronic Medical Records, and Electronic Health Records in Epidemiology in Japan: Narrative Literature Review
JMIR Med Inform 2023;11:e39876
URL: https://medinform.jmir.org/2023/1/e39876
doi:10.2196/39876
PMID:36787161

Review

Smart Glasses for Supporting Distributed Care Work: Systematic Review

Zhan Zhang[1], PhD; Enze Bai[1], MSc; Karen Joy[1], MSc; Partth Naressh Ghelaa[1], MSc; Kathleen Adelgais[2], MD; Mustafa Ozkaynak[3], PhD

[1]School of Computer Science and Information Systems, Pace University, New York, NY, United States

[2]School of Medicine, University of Colorado, Aurora, CO, United States

[3]College of Nursing, University of Colorado, Aurora, CO, United States

Corresponding Author:
Zhan Zhang, PhD
School of Computer Science and Information Systems
Pace University
1 Pace Plaza
New York, NY, 10078
United States
Phone: 1 3153992627
Email: zzhang@pace.edu

Abstract

Background: Over the past 2 decades, various desktop and mobile telemedicine systems have been developed to support communication and care coordination among distributed medical teams. However, in the hands-busy care environment, such technologies could become cumbersome because they require medical professionals to manually operate them. Smart glasses have been gaining momentum because of their advantages in enabling hands-free operation and see-what-I-see video-based consultation. Previous research has tested this novel technology in different health care settings.

Objective: The aim of this study was to review how smart glasses were designed, used, and evaluated as a telemedicine tool to support distributed care coordination and communication, as well as highlight the potential benefits and limitations regarding medical professionals' use of smart glasses in practice.

Methods: We conducted a literature search in 6 databases that cover research within both health care and computer science domains. We used the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) methodology to review articles. A total of 5865 articles were retrieved and screened by 3 researchers, with 21 (0.36%) articles included for in-depth analysis.

Results: All of the reviewed articles (21/21, 100%) used off-the-shelf smart glass device and videoconferencing software, which had a high level of technology readiness for real-world use and deployment in care settings. The common system features used and evaluated in these studies included video and audio streaming, annotation, augmented reality, and hands-free interactions. These studies focused on evaluating the technical feasibility, effectiveness, and user experience of smart glasses. Although the smart glass technology has demonstrated numerous benefits and high levels of user acceptance, the reviewed studies noted a variety of barriers to successful adoption of this novel technology in actual care settings, including technical limitations, human factors and ergonomics, privacy and security issues, and organizational challenges.

Conclusions: User-centered system design, improved hardware performance, and software reliability are needed to realize the potential of smart glasses. More research is needed to examine and evaluate medical professionals' needs, preferences, and perceptions, as well as elucidate how smart glasses affect the clinical workflow in complex care environments. Our findings inform the design, implementation, and evaluation of smart glasses that will improve organizational and patient outcomes.

(JMIR Med Inform 2023;11:e44161) doi:10.2196/44161

KEYWORDS

smart glass; care coordination; telemedicine; distributed teamwork; mobile phone

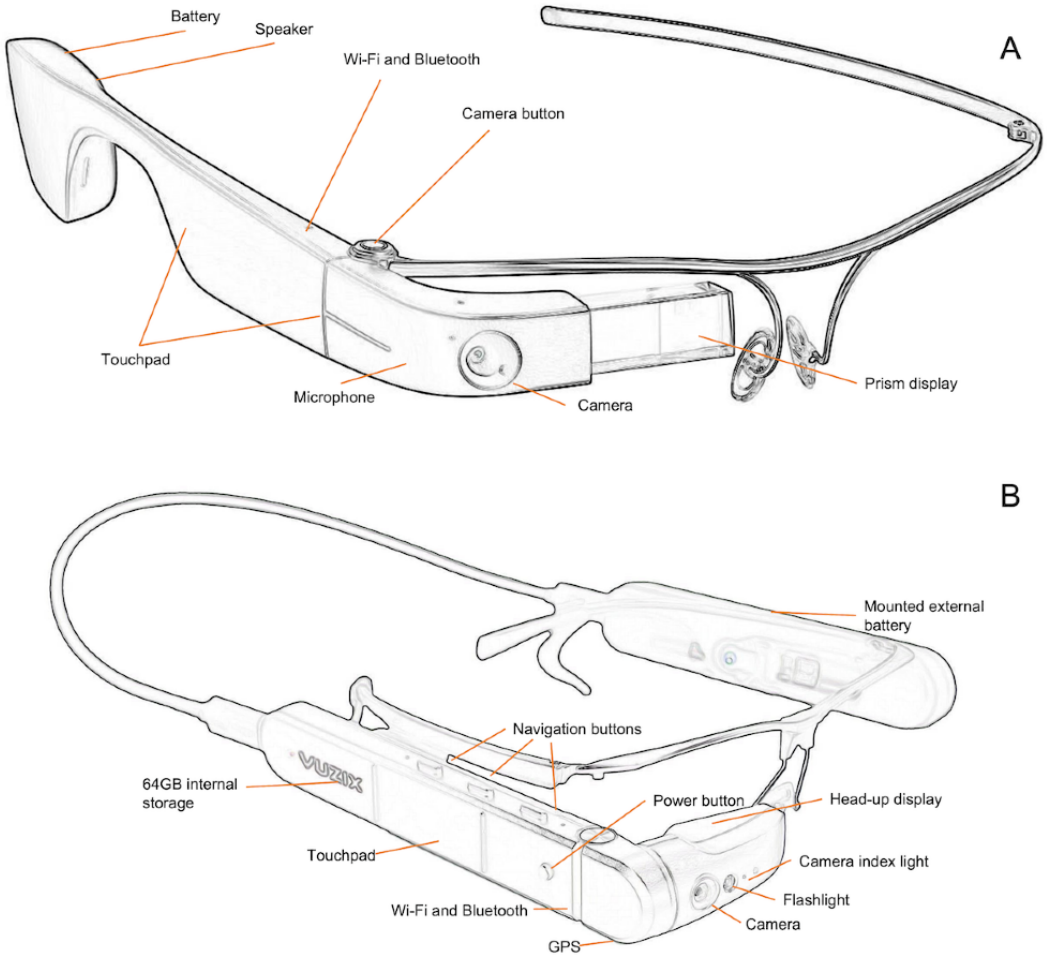
Introduction

Background

Effective and timely care coordination and communication are critical components of efficient and safe patient care [1,2]. Failure in providing coordinated care and communicating patient data is seen as one of the root causes of adverse events such as delays in patient care and deviations from standard medical procedures [3]. The challenges in maintaining effective care coordination and communication are exacerbated when care providers are distributed (eg, located in different places) [4,5]. Over the past 2 decades, many telemedicine systems have been developed to augment remote clinical consults [6-8]. During the COVID-19 pandemic, the need for such systems became more obvious. Most telemedicine systems are implemented on desktops or tablet devices [6,7]. However, these devices have practical limitations: (1) desktop systems have limited portability because they are installed in a fixed location; and (2) tablet device–based systems rely on manual input and control, which can hinder usability [9,10]. These issues could result in limited use of technology in real time, especially during complex care environments and time-critical patient scenarios because they demand the full cognitive attention and physical involvement of care providers [11].

In recent years, the use of smart glasses—a computing device worn as a conventional pair of glasses (Figure 1)—has been gaining momentum in health care because they allow for real-time visual communication in a hands-free manner [12,13]. In particular, smart glasses can present both imagery and textual information within the wearer's field of view (FOV) through a prism and enable videoconferencing for consults or second opinions via a front-facing camera. Since the introduction of smart glasses to the market, researchers have explored their applicability and usefulness in various medical settings and clinical scenarios [9], such as broadcasting surgeries to facilitate resident teaching [14], recording encounters with patients in wound care [15,16], assessing patients in mass casualty incidents [17], and supporting communication between prehospital and hospital providers [18,19].

Figure 1. Examples of smart glasses with various hardware components labeled. (A) Google Glass. (B) Vuzix M400.



Objectives

As there is a growing interest in using smart glasses to support care coordination and communication across distributed care providers [9,11,20], the aim of this study was to synthesize the knowledge and experiences in this area, understand the benefits and limitations regarding adopting smart glasses as a telemedicine tool, and inform the design of future smart glass applications to better support remote care coordination. We focused on the use of smart glasses in care coordination in various clinical settings (eg, surgical operation, emergency care,

and intensive care unit). Our specific research questions were as follows:

1. What are the general characteristics of prior research on using smart glasses for care coordination?
2. How was the system designed, used, integrated, and evaluated in supporting communication and care coordination across distributed care providers?
3. What types of challenges were identified by medical providers while they were using or testing the smart glass technology in practice?

These research questions were answered through a systematic literature review covering research within both health care and computer science fields.

Our work contributes the following to the medical informatics community: (1) an in-depth analysis and synthesis of prior research on the use of smart glasses for care coordination and communication; and (2) methodological and design implications for future research on smart glasses to improve distributed care coordination and communication.

Methods

Data Search

Our search started with discussing the search time frame and the most appropriate databases to use as well as search terms with experienced librarians. Using technology keywords such as "smart glasses" and "heads-up display," along with health care keywords such as "distributed care" and "telemedicine," a health librarian performed database searches for articles published between January 1, 2000, and March 1, 2022. We chose this time frame to capture the evolvement of this technology (ie, from early concepts such as head-worn displays [21] to smart glasses, which became a well-known concept after the introduction of Google Glass in 2013 [22]). The full list of search terms is presented in Textbox 1. We chose the following databases to cover research within both health care and computer science: ACM Digital Library, Cochrane Library, IEEE Xplore, Ovid MEDLINE, Embase, and Web of Science. A sample search strategy for Ovid MEDLINE is illustrated in Textbox 2. The database searches were set to include only studies published in peer-reviewed journals and conference proceedings in English. Literature reviews, dissertations, posters, and extended abstracts were excluded from the literature search. The retrieved citations were stored and managed using EndNote bibliographic management software (version X9; Clarivate).

Textbox 1. Keywords for literature search.

Search concepts and specific keywords

- Smart glass: *smart glass, augmented reality glasses, heads-up display, head-mounted, head-worn, virtual reality, augmented reality, mixed reality, wearable technology, Google Glass, Vuzix, Epson Moverio*
- Clinical: *distributed care, remote care, telehealth, telemedicine, telecare, emergency care, pre-hospital*

Textbox 2. A sample search strategy for MEDLINE.

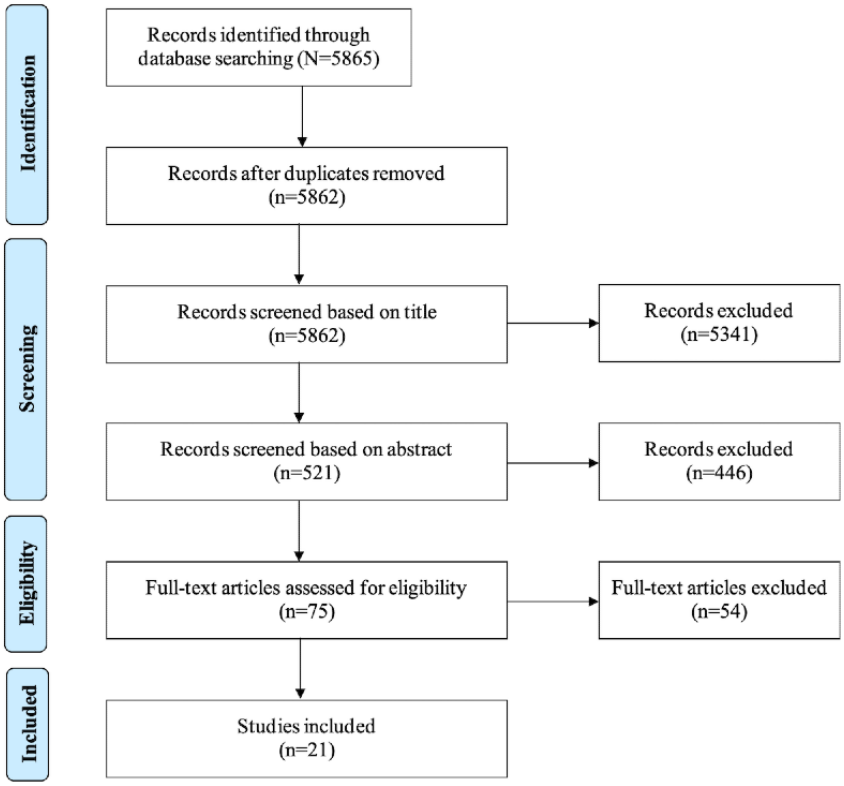
Search steps

1. ("distributed healthcare" or "distributed care" or "remote care" or tele* or nursing or "long term care" or "home health" or "home care" or prehospital or pre-hospital or "emergency medical" or "emergency care" or paramedic* or ((clinical or surg*) adj3 (application* or use* or implementation*))).ti,ab,kf. or exp Telemedicine/ or exp Home Care Services/ or exp Emergency Medical Services/
2. ((smart adj1 glass*) or smartglass* or Hololens or picolinker or (google adj1 glass*) or vuzix or "epson moverio" or "augmented reality" or (AR and augmented) or "mixed reality" or "virtual reality" or (VR and virtual) or "wearable technology" or wearables or "heads up" or "head mounted" or "head worn").ti,ab,kf. or wearable electronic devices/ or smart glasses/ or augmented reality/ or virtual reality/
3. Steps 1 and 2
4. Limit step 3 to (english language and yr="2000-Current")
5. (training* or education* or simulation* or telephon* or teleconferenc* or television*).ti. or exp *education/ or *telephone/ or *television/
6. Step 4 not step 5

Article Screening and Selection

We used the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) methodology to search and screen articles [23]. Figure 2 outlines the number of records that were identified, included, and excluded through different phases. More specifically, 5865 articles were identified through database searches, of which 5862 (99.95%) were included for screening after removing duplicates. Article titles were screened first, followed by abstract screening, to identify relevant articles. Of the 5862 articles, after screening of article titles, we excluded 5341 (91.11%); of the remaining 521 studies, 446 (85.6%), were excluded, leaving 75 (14.4%) for full-text review. After reviewing the full text of these 75 articles, we deemed 21 (28%) to be eligible for this systematic review.

Figure 2. Information source and search strategy.



Three authors (EB, KJ, and PG) independently screened all papers through the paper stack and selected relevant papers for inclusion. Two senior researchers (ZZ and MO) oversaw the whole article review and selection process. Any conflicts in selection decisions were resolved through discussion among all the authors during weekly group research meetings. The inclusion criteria were peer-reviewed articles that reported the use or testing of any smart glass technology and accompanying software in the context of communication and collaboration across distributed care providers. Articles were excluded if they only reported the use of smart glasses by an individual or in a collocated clinical setting or if they did not provide adequate supporting information, such as what clinical setting the smart glasses were used in and who used the technology.

Data Extraction, Analysis, and Synthesis

Guided by the research questions of this study, 2 authors (KJ and EB) used a Microsoft Excel spreadsheet to extract, collate, and summarize data from the included studies, such as the country where the study was conducted, study objectives and scope, clinical scenarios, system evaluation methods, technology specifics, barriers and challenges, and a summary of study findings. Textbox 3 summarizes these data fields and their brief definitions. In addition to extracting the aforementioned metadata, we also assessed the technology readiness levels (TRLs) [24] of the systems tested in the reviewed studies. There are 9 different TRLs, ranging from level 1 (scientific knowledge generated underpinning hardware and software technology) to level 9 (actual system "flight proven" through successful mission operations). Two authors (KJ and EB) followed the metrics proposed in the study by Engel et al [25] and independently assessed TRLs for each system. They then compared and discussed their TRL evaluations until they reached agreement.

Two senior researchers (ZZ and MO) reviewed all the articles and analyses as a verification step. The research team met regularly to discuss the results. We performed the data analysis iteratively (ie, we went back and forth as more knowledge was obtained), as suggested by prior work [11,26]. A meta-analysis of the study results was not considered in this work owing to the heterogeneity of the study designs and results.

In the following section, we report information that was synthesized from the reviewed articles, including characteristics of the selected studies, system architecture and features, TRLs of the reviewed systems, system evaluation methods, and care providers' perceived benefits and challenges of using and adopting smart glasses for distributed care coordination.

Textbox 3. Assessed article information and metadata.

Assessed information and brief definition

- Study objectives and scope: the objective of the research and the purpose and scope of the use and test of smart glasses in each study (eg, patient care vs medical training)
- Clinical scenarios: the clinical domain and context in which the study was conducted
- Publication details: the type (eg, journal article vs conference paper), region, and year of the publication
- System infrastructure: the hardware, software, and network setup on both local and remote sites for establishing teleconsultation
- System features: the system features used, developed, or evaluated in each study
- System evaluation: the aspects of the smart glass system that were evaluated in the study and the methods used for system evaluation
- Benefits and challenges: the reported benefits and challenges of using smart glasses in improving communication and care coordination among distributed medical teams
- Major study findings: a summary of the major findings of a study

Results

General Characteristics of the Reviewed Studies

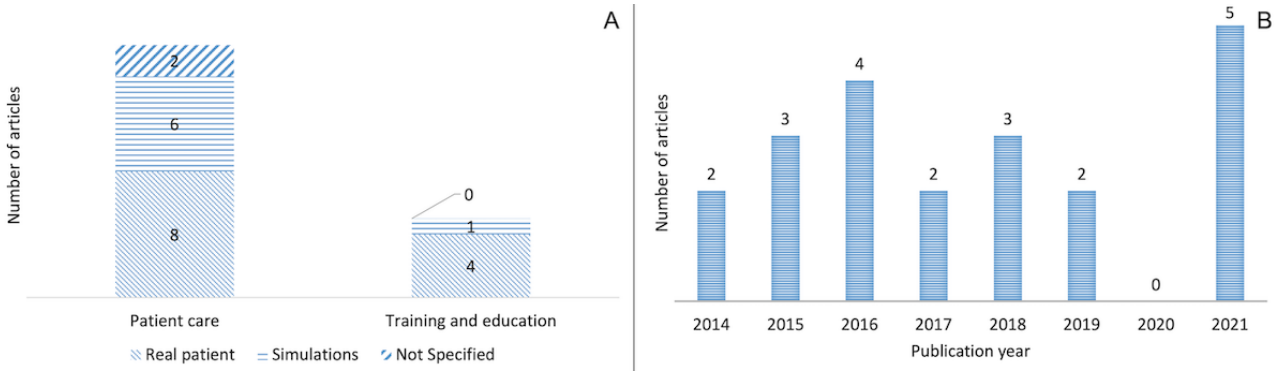
Of the 21 reviewed articles, 10 (48%) were conducted in the United States [18,19,27-34], and 2 (10%) were conducted for surgical teleproctoring between high-income countries and low- and middle-income countries (LMICs), such as between surgeons in the United States and Mozambique [35] and between experienced surgeons recruited from the United States and Germany and novice surgeons in Brazil and Paraguay [36]. The remaining studies (9/21, 43%) were conducted in different countries, such as Spain [37], China [38], Germany [39], France [40], Italy [41], Switzerland [42], Malaysia [43], South Korea [44], and Republic of the Congo [45]. The reviewed studies were conducted to assess the feasibility, effectiveness, and user experience of smart glasses in supporting remote patient evaluation and care procedure operation in a particular medical domain. The study objectives, along with major findings for each reviewed article, are presented in Multimedia Appendix 1 [18,19,27-45].

The clinical foci in these 21 papers vary: 9 (43%) focused on surgical settings [29,30,33-38,44], whereas 6 (29%) focused on the prehospital or emergency medical services domain [18,19,28,31,39,42]. The remaining studies (6/21, 29%) focused on intensive care [40,43], toxicology [27], ophthalmology [32], pediatric cardiology [41], and general medicine [45].

The scope and purpose of the use of smart glasses among these studies vary. As shown in Figure 3A, the majority of the reviewed studies (16/21, 76%) used smart glasses to enable remote patient care and evaluation [18,19,27,28,30-32,37-45]. Of these 16 studies, 8 (50%) [27,28,30,32,37,38,43,45] tested smart glasses with real patients, 6 (38%) [18,19,31,39,40,44] conducted system testing in a simulated environment, and 2 (13%) [41,42] did not specify how the device was tested. The remaining studies (5/21, 24%) [29,33-36] leveraged smart glasses for training and teleproctoring purposes; of these 5 studies, 4 (80%) [29,34-36] tested the device with real patients, whereas 1 (20%) [33] tested the device in a simulated environment.

The reviewed articles were published between 2014 and 2021 (Figure 3B). It is noticeable that almost half of the reviewed articles (9/21, 43%) were published within the first 3 years of the release of Google Glass [22]. Subsequently, the number of studies on the use of smart glasses for supporting distributed care decreased until 2021. One possible explanation for this finding is that the use of smart glasses regained momentum right after the outbreak of the COVID-19 pandemic as researchers started exploring smart glass use to enable medical personnel to participate in remote assessment and consultation, with the aim of safeguarding patients and health care providers during the pandemic.

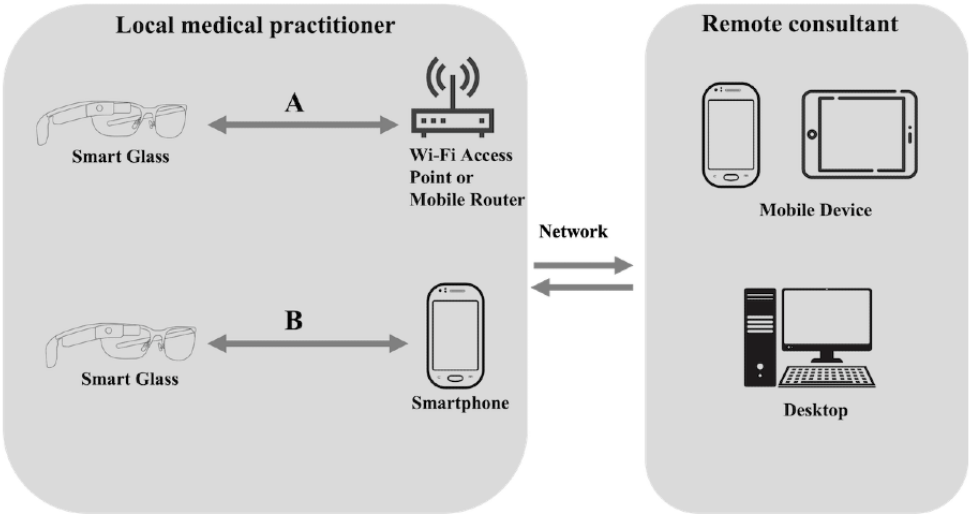
Figure 3. (A) The scope and testing environment of smart glasses in the reviewed articles. (B) The distribution of reviewed articles over the years.



System Architecture

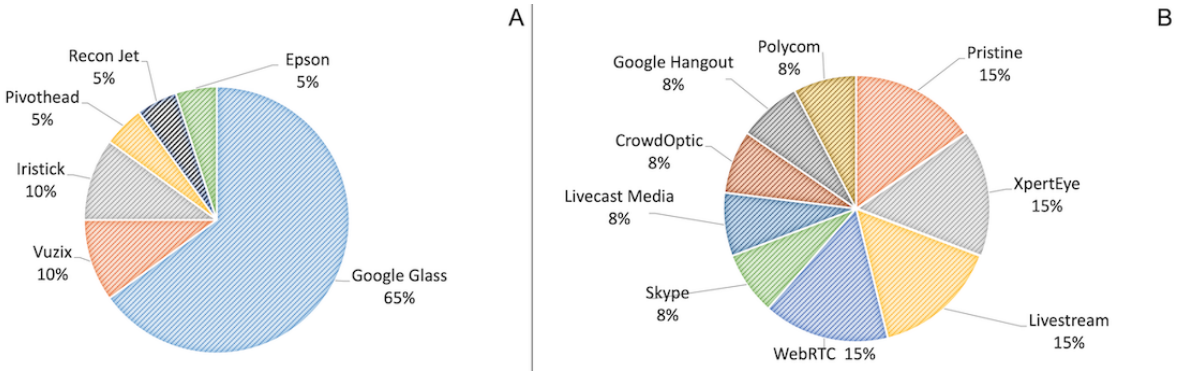
Although the system architecture implemented in each study varied, there were some similarities across the reviewed studies. Typically, there are two types of technology setups on the local site: (1) smart glasses are connected to a Wi-Fi network, a Wi-Fi hotspot, or a mobile router to directly stream the first-person point-of-view to a remote consultant (Figure 4A); or (2) smart glasses are connected to a smartphone or a laptop via Bluetooth or Wi-Fi for video streaming and audio transmission (Figure 4B). The first approach was adopted by 52% (11/21) of the studies [19,27,28,31,32,36,38,39,41,43,44], and the second approach was used in 33% (7/21) of the studies [18,29,35,37,40,42,45]; for example, in the study by Diaka et al [45], the smart glasses were designed as an extension of a smartphone, which meant that the local wearer needed to initiate the call on the smartphone. Regardless of the system implementation method on the local site, the remote experts were usually equipped with either a computer or a mobile device (eg, a tablet device) to review and access the video stream and other multimedia data shared by the local medical practitioner (Figure 4). However, it is worth mentioning that in the study by Brewer et al [33], where smart glasses were used for surgical training, the remote expert (trainer) also wore a pair of smart glasses to view the video streamed from the learner.

Figure 4. Common system architecture setups in the reviewed studies. (A) Smart glasses connected to a Wi-Fi network, a Wi-Fi hotspot, or a mobile router. (B) Smart glasses connected to a smartphone via Bluetooth or Wi-Fi.



As shown in Figure 5A, the reported brands of smart glasses in these studies included Google Glass [18,19,27-30,33-36,40,42,44], Vuzix [38,43], Iristick [37,45], Pivothead Original Series [32], Intel Recon Jet [31], and Epson Moverio BT-200 [41]. Google Glass was the most frequently used smart glass device (13/20, 65%). Another interesting observation is that all of the studies (21/21, 100%) used off-the-shelf, commercialized videoconferencing software (Figure 5B) such as Pristine Eyesight [19,27], AMA XpertEye [28,35], Livestream [18,36], WebRTC (enabled by Google) [42,44], Livecast Media [38], Skype [29], CrowdOptic [33], Google Hangout [34], and Polycom RealPresence Group 500 [32]. Most of the videoconferencing software used was compliant with the Health Insurance Portability and Accountability Act (HIPAA) rules, except in the case of the study by Cicero et al [18], where the researchers only tested the use of smart glasses in a simulated environment (real patient care was not involved).

Figure 5. (A) Smart glass brands used in the reviewed articles. (B) Videoconferencing tools used in the reviewed articles.



System Features

Although there was variation in the application scopes and domains, there were some common software features across the reviewed studies (Textbox 4). Real-time synchronous video and audio streaming from the local smart glass wearer to the remote consultant is the most common feature among the studies (19/21, 90%). In the case of the exceptions (2/21, 10%), because of technical limitations (eg, limited internet connection), the study by Gupta et al [30] first recorded patient care and evaluation

using smart glasses and then transmitted the recordings to remote experts at a later time to simulate real-time telemedicine consults, whereas in the study by Hashimoto et al [34], researchers used Google Glass and an Apple iPhone to capture videos of a surgical operation and compared the video quality and its adequacy for safe use in telementoring.

Another noteworthy feature is enabling imagery and text-based remote guidance and annotation; for example, the remote consultant can annotate images captured from the live stream and project them back onto the local glass wearer's visual field [35,37]. In 19% (4/21) of the studies [19,27,36,44], the remote consultant could use the texting feature to type messages that could be projected onto the smart glass display. These annotation features provide the remote consultant with more channels (in addition to audio and video) to direct and guide local medical practitioners to perform critical procedures.

Augmented reality (AR)—a technique that can enhance an individual's visual experience of the real world through the integration of digital visual elements—was also tested in several studies. In Ponce et al [29], for example, AR enabled a remote surgeon to insert their hands or instruments virtually into the visual field of the local surgeon who wore smart glasses for real-time guidance, training, and assistance as needed. In another study [41], a remote specialist used AR-based markers to guide the execution of an echocardiographic examination performed by a local operator. The markers were overlaid on the ultrasound device and could be seen through the screen of the local operator's smart glasses.

Other features of smart glasses reported in the studies included zooming in and out of the live stream video [35]; using voice commands [27,28,30,31] or head movements [27] to control, and interact with, the smart glass device; taking photographs [19,30,31,35]; automatically detecting the geographic location of on-site medical teams with the built-in GPS [31]; and presenting prehospital triage algorithm on the glass screen for decision support during mass casualty incidents [39].

Textbox 4. Summary of smart glass features as described in the reviewed studies.

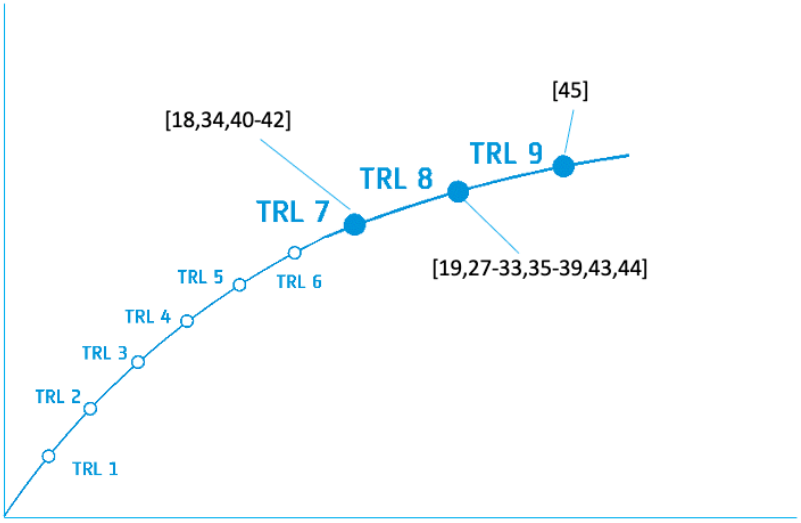
System features

- Real-time synchronous video and audio streaming [18,19,27-29,32-38,40-45]
- Record and forward video recordings [30]
- Imagery and text-based remote guidance and annotation [19,27,35-37,44]
- Augmented reality [29,41]
- Zooming in and out of the live stream video [35]
- Hands-free interaction with smart glasses [27,28,30,31]
- Taking photographs [19,30,31,35]
- GPS-based tracking of the geographic location of on-site medical teams [31]
- Presenting prehospital triage algorithm on the glass screen for decision support [39]

TRLs of the Systems Tested in the Reviewed Studies

On the basis of our analysis, we found that the TRLs of all the systems used or tested in the reviewed studies ranged between 7 and 9. Our TRL assessment for each system is visualized in Figure 6 [18,19,27-45]. The reasoning for our assessment is summarized in Multimedia Appendix 2 [18,19,27-45].

Figure 6. Diagram of technology readiness levels (TRLs) for the systems reported in the reviewed studies [18,19,27-45].



The systems in 24% (5/21) of the studies [18,34,40-42] have a TRL of 7, which indicates that the technology is in the form of a high-fidelity prototype and has all key functionality available for demonstration and test; for example, in the study by Widmer and Müller [42], the Google Glass device on the local site was set up to connect with a computer application on the remote site for teleconsultation. This integrated system was only preliminarily tested by the research team but not in a simulated or real environment (a criterion for TRL 8); thus, its TRL was set to 7. It is worth mentioning that of these 5 studies, 3 (60%) [18,40,41] tested smart glasses in simulated scenarios; however, there were several reasons for their failure to meet the criteria for TRL 8, such as using non–HIPAA-compliant videoconferencing software, testing the technology with only 1 volunteer, or not fully integrating smart glasses with the network and remote devices.

The majority of the studies (15/21, 71%) [19,27-33,35-39,43,44] tested or used systems that met the criteria for TRL 8, indicating that they are actual systems in their final configuration and have been fully developed and tested in either simulated or real operational scenarios. However, these studies provided limited information regarding some criteria for TRL 9, such as whether the system had been fully integrated with other operational hardware and software systems (eg, database and hospital IT infrastructure), whether all system documentation had been completed, whether training on system use was available, and whether engineering support team was in place. Without such information, it is difficult to assess the readiness of these systems for large-scale deployment.

In comparison, only the system in the study by Diaka et al [45] was assessed to have a TRL of 9 because the system had been successfully operated on actual missions and tasks in the operational environment for a relatively long time (ie, more than a year). Furthermore, the system was fully integrated with other operational software, hardware, and network devices, as well as care delivery services (eg, moto-ambulances to facilitate patient referrals after teleconsultation).

System Evaluations

Overview

The reviewed studies evaluated different dimensions of the smart glass system, including technical feasibility, effectiveness, and user experience. The details regarding the aspects of the smart glass system that were evaluated as well as the evaluation methods used in the reviewed studies are summarized in Table 1 and then elaborated on in the following sections.

Table 1. Summary of system evaluation details.

Evaluated dimensions	Specific evaluated aspects	Evaluation methods
Technical feasibility [27,34-36,44]	• Success rate of established video teleconsultations between local and remote medical practitioners [27,36] • Whether the quality of video and audio streaming was good enough for enabling video streaming [34-36,44]	• Researchers' observations of the successfulness of teleconsultations [27,36] • Questionnaire [27,34-36,44]
Effectiveness [18,19,27,28,30-33,36,39,40,43]	• Compared with in-person patient evaluation, whether the use of smart glasses could achieve similar performance and accuracy regarding patient evaluation and diagnosis [19,28,32,43] • Compared with either mobile phone–based or no remote patient consultation, whether the use of smart glasses could lead to changes in clinical management and remote consultant's confidence regarding diagnosis [18,27,30,39,40] • Whether the use of smart glasses could improve medical training (eg, surgical operation) [33,36]	• Comparison study between control (without smart glass support) and treatment (with smart glass support) groups [18,19,28,32,39,40,43] • Questionnaire [33] • Exit interview [36]
User experience [18,19,27,29-31,35,37-41,43-45]	• Usability of smart glasses [19] • Opinions regarding using and adopting smart glasses in practice [18,27,30,31,35,37-41,43-45]	• Survey [18,19,27,29-31,35,38-40,43,44] • Interviews and observations [18,31,35,45]

Technical Feasibility

Several studies assessed whether the smart glass technology was a practical means to support care coordination and communication in different contexts, such as teletoxicology consults [27] and remote surgical teleproctoring [34-36,44]. The main measurements included the success rate of established video teleconsultations between local and remote medical practitioners and whether the quality of video streaming was acceptable and good enough to allow for real-time, seamless guidance and assistance. The technical feasibility was primarily determined by the researchers' observations and the users' ratings via questionnaire; for example, in a study evaluating the feasibility and acceptability of Google Glass for teletoxicology consults [27], questionnaires were administered immediately after the study to elicit remote consultants' opinions regarding whether consults through smart glasses were considered successful and the technical feasibility of using smart glasses for teleconsultation.

Effectiveness

Of the 21 reviewed studies, 10 (48%) evaluated the effectiveness of smart glasses, that is, whether this novel technology could improve patient care and decision-making compared with current approaches (eg, no remote consultation, in-person patient evaluation, or consultation via telephone)

[18,19,27,28,30,32,33,39,40,43]; for example, in some settings where remote consultations were usually accomplished via telephone or radio, which typically do not support visual communications [27,30], researchers compared using such traditional communication mechanisms with using smart glasses to determine whether the use of smart glasses could lead to changes in clinical management and the remote consultant's confidence regarding diagnosis.

Of these 10 studies, 7 (70%) [18,19,28,32,39,40,43] conducted an experiment with a control group (no smart glasses and either in-person consultation or no remote consultation at all) and an intervention group (with smart glasses) to measure whether using smart glasses could increase the quality and accuracy of patient diagnosis while reducing the time needed to perform patient care; for example, in the scenario of patient triage during mass casualty incidents [19], researchers asked 2 emergency medicine (EM) physicians (control group) to make triage decisions after examining the simulated patients in person as 2 other EM physicians (intervention group) simultaneously evaluated the same group of patients via real-time point-of-view video stream from a paramedic wearing Google Glass. They then used the agreement within and among the groups of EM physicians on the need for immediate trauma evaluation to determine the effectiveness of smart glasses for supporting patient triage.

User Experience

Of the 21 studies, 15 (71%) examined end users' experience and perceptions to some extent with regard to using smart glasses in their work [18,19,27,29-31,35,37-41,43-45]. The primary methodology used for eliciting user experience was a survey, which was adopted by 80% (12/15) of these studies [18,19,27,29-31,35,38-40,43,44]; for example, in a recent study [43], a survey was sent to the participants on completion of the study to assess acceptance, satisfaction, overall impact, efficacy, and potential of adopting smart glasses as an alternative method of teleconsultation in neurosurgery. Among these 12 studies that administered a survey, 9 (75%) specifically reported the number of participants, which ranged between 2 and 276. Other methods such as interviews and observations were also used to gather more qualitative, in-depth insights from end users [18,31,35,45]. In particular, of these 4 studies, 2 (50%) [31,35] conducted interviews in conjunction with a survey.

It is also worth mentioning that of the 15 studies, 2 (13%) [19,31] specifically focused on evaluating the usability of smart glasses, that is, whether smart glass technology is perceived as easily usable by, and acceptable to, medical professionals. Another study [30] also examined patient perceptions of medical providers wearing smart glasses with recording capability. Finally, of the 15 studies, 5 (33%) [29,37,38,41,42] mentioned that they collected end users' opinions and experiences but did not specify the methods they used.

Benefits and Challenges of Using and Adopting Smart Glasses for Teleconsultation

Benefits

Our reviewed work highlights the advantages of smart glasses in improving communication and care coordination among distributed medical teams because this technology enables local medical providers to share visual information and perform teleconsultation in a hands-free manner. Regarding the effects on clinical care and patient outcome, the studies reported that smart glasses could shape clinical management and boost remote consultants' confidence in clinical care [27,30], achieve diagnostic accuracy comparable with that achieved in in-person patient examination [19,28,32,43], improve proficiency and performance of the clinical tasks [31,33,35,38-40], and lower the medical service cost and improve quality of life for people in rural areas or LMICs [36,38]. Finally, many studies reported positive user perceptions, acceptance, and satisfaction with the use of smart glasses [19,27,29-31,35,38,39,41,43,45].

Notwithstanding these reported benefits, the reviewed studies also highlight a set of challenges and user concerns regarding the adoption of smart glasses in practice. We grouped them into 4 main categories: technical challenges, human factors and ergonomics, privacy and security concerns, and organizational challenges (Textbox 5).

Textbox 5. Challenges to using and adopting smart glasses in practice.

Technical challenges

- Unstable or low-bandwidth internet connections [18,19,29,33,35,36,39,44]
- Battery drain becomes higher during video streaming [18,29,39]
- The microphone is unable to filter out background noise [18,29]
- Screen contrast and readability issues in bright or dark environments [18]
- Image distortion owing to overexposure to room light [18,29,35]
- Smart glass see-through screen is too small for easy interaction [41]
- Difficulty controlling video streaming software [18,35,38]
- Lack of a lock function to prevent the possibility of inadvertently halting the video streaming and ability to opt out of frequent software updates [18]

Human factors and ergonomics

- Compatibility issues with wearer's glasses or personal protective equipment [27,29,35,37,39-41]
- Misalignment between the direction of gaze and range of smart glass camera [29,35,37,40,41,43]
- Voice control function could be problematic [18,30]
- Added distractions for medical professionals [31]

Privacy and security concerns

- Concerns regarding violations of patient privacy and data breach [28-30,43]

Organizational challenges

- Added workload for medical professionals [39]
- Costly device and software [35]
- End users have limited experience with, and prior knowledge of, smart glasses; need extensive equipment and software training [27,37,41,43]

Technical Challenges

The reviewed studies reported a variety of technical challenges that may impede the effective use of smart glasses in teleconsultation. These challenges are mainly related to internet connections, hardware limitations, and software reliability. More specifically, because smart glasses require a high-speed network to transmit visual media (eg, video streaming, audio, and pictures), unstable or low-bandwidth internet connections were seen as a major technical barrier because this issue would compromise video and audio quality, leading to breakdowns in communication and loss of patient information [18,19,29,33,35,36,39,44]. This is more evident in low-resource or out-of-hospital settings where medical practitioners have limited access to the internet; for example, because Wi-Fi is not steadily available in the prehospital environment, the problem with internet connections was commonly reported in this domain [18,19,39]. One practical and successful solution used by a study in prehospital communication [31] was using a mobile router to provide a fault-tolerant network that ran independent of Wi-Fi and other external networks, allowing for deployment at any location.

Regarding hardware limitations, medical professionals were concerned about battery life (eg, the battery could get drained quickly during video streaming) [18,29,39], microphone sensibility (eg, not being able to filter out background noise) [18,29], screen contrast and readability (eg, hard to read the screen in extremely bright or dark environment) [18], image quality (eg, the image could be distorted because of overexposure to room light) [18,29,35], and small screen for interaction [33,41,44].

Issues regarding software were primarily related to controlling and interacting with the video streaming software; for example, 14% (3/21) of the studies [18,35,38] mentioned difficulties regarding zooming in or out during video streaming; as such, the smart glass wearer needs to bring their face close to the patient. Other software issues included the lack of a lock function to prevent the possibility of inadvertently halting the video streaming and the inability to opt out of frequent software updates [18].

Human Factors and Ergonomics

Many issues related to the interactions between users and the smart glass system were also reported. First, 38% (8/21) of the studies [27,29,33,35,37,39-41] highlighted the compatibility issue with users' spectacles or personal protective equipment. In particular, fitting the smart glass headset onto surgical loupes was problematic, interfering with the surgeon's ability to wear such devices [35]. Some users had to remove their spectacles to wear the smart glass headset or tie up their hair to prevent the glass camera from being hidden [40]. Second, the difference in line of sight—misalignment between what the glass wearer sees (eg, the direction of gaze) and what the camera captures (eg, range and angle of the camera)—was also cited as a major

barrier [29,35,37,40,41,43]. This issue was often attributed to the limited FOV of smart glasses [33,44]. This misalignment problem could be worsened owing to sudden head movements and frequent relocation of the smart glass wearer or the patient's unpredictable movements because these could cause motion blur for remote experts or consultants and make it difficult for them to identify the clinical situation [44]. Third, although the reviewed studies reported that their participants perceived that the smart glass was easy to use overall, usability issues still exist; for example, the voice control function did not work perfectly and thus required the user to remove their gloves to use the built-in touchpad or buttons to operate the device, such as starting or stopping the video call [18,30]. In another study, smart glasses were reported to be a distraction for medical practitioners [31].

Privacy and Security Concerns

Patient privacy and data security issues were perceived as important to address because smart glasses can transfer or even store sensitive patient data [28-30,43]. These studies stated that any implementation of smart glasses must not only comply with HIPAA requirements but also alleviate patient concerns about any potential privacy violation or misuse of their data [30,43].

Organizational Challenges

As medical professionals have limited prior knowledge of using the novel smart glass technology (compared with their experience of using smartphones or tablet devices), a few studies mentioned that user training is necessary to increase efficiency and reduce human errors in system operation [27,33,37,41,43]. In addition, the smart glass technology is costly; for example, as McCullough et al [35] reported, the cost of a yearly contract for a piece of wearable hardware and the videoconferencing platform is approximately US $7000. Such high costs could become a critical barrier to adopting this technology at scale, especially for those health care providers who have limited resources. Finally, integrating smart glasses into the current workflow is a prominent challenge; for example, Follmann et al [39] reported that adopting smart glasses in prehospital triage and communication added more workload to emergency care providers in the field and took markedly more time compared with not using smart glasses.

Discussion

Methodological Implications

In this work, we conducted a systematic review of studies focused on the use and application of smart glasses in supporting care coordination and communication among distributed medical teams. Of the 5862 papers included for screening, only 21 (0.36%) met our criteria, highlighting the paucity of studies examining the feasibility, effectiveness, and user experience of using smart glasses as a telemedicine tool. Furthermore, the studies were mostly conducted in the United States and a few other high-income countries (eg, Italy, Germany, and France). One possible explanation is that smart glass technology is costly, hindering its adoption in LMICs and low-resource settings. However, 14% (3/21) of the reviewed studies [35,36,45] revealed the substantial benefits that smart glasses could bring to LMICs and rural areas, such as providing remote training and mentoring and more accurate instructions to the field medical practitioners in low-resource settings who otherwise have limited access to remote experts. Given such benefits, more future work is needed to expand the research of smart glasses to LMICs.

Another interesting observation is that all the reviewed studies (21/21, 100%) only used off-the-shelf hardware and software without involving users in the system design process. Prior work has suggested that it is critical to involve users and understand user requirements in the early phase of system development to identify and address potential usability and technical issues [6,46,47]. In addition, regarding the methodology for eliciting user opinions, out of 15 studies conducted user evaluation, 33% (5/15) of them did not specify what questions they asked, how the questionnaire was developed, and what procedure was followed. Despite the user-friendliness of health care information technology being a determinant factor for user adoption and acceptance [48,49], the usability of smart glasses was neglected by most of the studies (19/21, 90%), with only the studies by Broach et al [19] and Demir et al [31] specifically examining this aspect. These facts highlight the need to adopt a *user-centered design* approach in the development of smart glass technology by placing users at the center of the system design process from inception to implementation and deployment.

A similar concern is that a few of the reviewed studies (4/21, 19%) only recruited a small number of study participants (eg, 2 health care professionals) to participate in their user studies (eg, survey or interview). In addition, some of the studies (5/21, 24%) did not report the details of their user research, including the number of participants. These findings may suggest that the important role of user research was not recognized in some of the reviewed studies (9/21, 43%), and their results might not be generalizable because of the limited number of study participants. Given these study limitations, we argue that involving human-computer interaction researchers in such type of research and establishing close collaborations between these researchers and health care domain experts are critical and much needed, as demonstrated in the study by Schlosser et al [50].

Finally, almost all of the reviewed studies (20/21, 95%) focused on evaluating the smart glass technology either from a technical perspective or a clinical perspective, while neglecting other important factors that could substantially affect the use and adoption of this technology, such as workflow, teamwork, policies, and organizational cultures. As prior work has argued [51], an ongoing challenge to the successful implementation and deployment of health IT (HIT) interventions is to operationalize their use within the workflow of a complex health care system; for example, a new technology could disrupt current clinical work, causing not only frustrations for medical providers but also patient safety issues [52-54]. When this problem occurs, not surprisingly, medical practitioners are left with no choice but to bypass the technology or adopt informal, low-tech, potentially unsafe workarounds that deviate from the formal protocol [55,56]. As such, researchers have highlighted the importance of examining the design, use, and application of HIT interventions through the lens of a sociotechnical

perspective [55-57]. This approach allows researchers and practitioners to understand the complex interrelations between various social and technical elements of systems that are equally important in determining the success of HIT adoption in a health care organization. In line with this argument, we believe that more research adopting a sociotechnical model [51,58] is needed to investigate the factors (eg, human-computer interaction, workflow and communication, internal organizational features, and external rules) that contribute to the uptake of smart glasses in routine use.

Design Implications

The reviewed studies revealed a set of challenges and barriers to adopting and using smart glasses in practice; for example, a commonly cited technical challenge is internet connection quality—smart glasses rely on a high-bandwidth internet network for streaming videos and transmitting other visual media data (eg, high-resolution pictures, texts, and augmented objects). However, this technical requirement could be challenging to fulfill, especially in low-resource or out-of-hospital settings [59]. With the rapid development of 5G technology, this technical barrier might be overcome in the near future; for example, a study [60] showed that 5G technology could not only enable safe and efficient complex surgical procedures during telementored surgery but also lead to a very high degree of surgical team satisfaction. In addition to internet connections, other technical improvements suggested by the reviewed studies include increasing the memory space of smart glasses to store more information, adding autofocus and stabilization features to the smart glass camera, and improving the camera resolution [35].

Human factors and usability issues make up another set of important considerations for smart glass designers and developers; for example, the difference in line of sight between the local medical practitioner and remote consultant impeded the remote consultant from seeing exactly what the smart glass wearer's eyes were fixed on. In addition, the limited FOV further complicated the video transmission to the remote experts. One reviewed study [44] experimented by attaching a mirror to the smart glass to increase the FOV of the local practitioner by transmitting both the wearer's front view and their hand operations below the camera to the remote experts. However, the video received on the other end by the experts was deemed confusing. Another viable solution suggested by prior work [59] is using more advanced mounting techniques to make sure that the smart glass can sit steadily on the wearer's head to align their visual field with the camera range. Another interesting issue brought out by a few of the studies (5/21, 24%) was the necessity of enhancing user interactions with the smart glass, such as offering more hands-free interaction mechanisms (eg, using head movements to control the device) [35] and enabling the user to zoom in and out during video streaming as well as pan the image [38].

Current smart glass applications are stand-alone and limit their potential. The data collected and transferred through smart glasses can best benefit patient care tasks if they can be incorporated into, and fully integrated with, other HITs such as electronic health records or clinical decision support systems. Interoperability issues (eg, standardized terminology) should be considered when deploying and integrating smart glasses into complex health care systems.

Other important design considerations that need full attention for developing and deploying the smart glass technology include (1) ensuring that the software is compliant with HIPAA requirements to protect patient privacy and data security, (2) integrating smart glasses into the workflow to minimize the disruption to medical practitioners' work, and (3) providing sufficient training to end users.

Study Limitations

Defining the search keywords was difficult. To generate a comprehensive and relevant list of keywords, we iteratively discussed and selected the keywords for the search based on suggestions from the health librarian and a review of systematic review articles regarding smart glasses. Another limitation is that we did not assess the quality or impact of the results from the included articles. A meta-analysis was not feasible because of the heterogeneity of the study designs and results.

Conclusions

Smart glasses were found to be an acceptable and feasible tool in enabling visual communication and information sharing among distributed medical teams. Despite the high potential of this novel technology, the reviewed articles pointed out a set of challenges that need to be addressed before the wide deployment of this technology in complex health care systems. Thoughtful system design involving end users from the beginning and improved hardware and software reliability are needed to improve the usefulness and usability of smart glasses for medical practitioners [11,59]. We suggest that more user-centered design and evaluation research is needed to examine and evaluate medical professionals' needs and perceptions and determine how to design smart glass technology to meet their needs. In addition, more research is required to elucidate how smart glasses affect the workflow of medical professionals in complex care environments.

Acknowledgments

The authors thank the health librarians Lilian Hoffecker and Ben Harnke at the University of Colorado for performing the literature search and documenting the search process and results. This study was supported by the National Science Foundation (grant 1948292) and the Agency for Healthcare Research and Quality (grant 1R21HS028104-01A1).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Summary of study objectives and major findings.
[DOCX File , 28 KB - medinform_v11i1e44161_app1.docx]

Multimedia Appendix 2
Technology readiness levels of the systems reported in the reviewed studies.
[DOCX File , 27 KB - medinform_v11i1e44161_app2.docx]

References

1. McDonald KM, Sundaram V, Bravata DM, Lewis R, Lin N, Kraft SA, et al. Closing the Quality Gap: A Critical Analysis of Quality Improvement Strategies (Vol. 7: Care Coordination). Rockville, MD, USA: Agency for Healthcare Research and Quality; 2007.
2. Schultz EM, McDonald KM. What is care coordination? Int J Care Coord 2014 Aug 27;17(1-2):5-24. [doi: 10.1177/2053435414540615]
3. Schultz EM, Pineda N, Lonhart J, Davies SM, McDonald KM. A systematic review of the care coordination measurement landscape. BMC Health Serv Res 2013 Mar 28;13:119 [FREE Full text] [doi: 10.1186/1472-6963-13-119] [Medline: 23537350]
4. Amir O, Grosz BJ, Gajos KZ, Swenson SM, Sanders LM. From care plans to care coordination: opportunities for computer support of teamwork in complex healthcare. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. 2015 Apr Presented at: CHI '15; April 18-23, 2015; Seoul, South Korea p. 1419-1428. [doi: 10.1145/2702123.2702320]
5. Zhang Z, Sarcevic A, Bossen C. Constructing common information spaces across distributed emergency medical teams. In: Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing. 2017 Feb Presented at: CSCW '17; February 25-March 1, 2017; Portland, OR, USA p. 934-947. [doi: 10.1145/2998181.2998328]
6. Zhang Z, Brazil J, Ozkaynak M, Desanto K. Evaluative research of technologies for prehospital communication and coordination: a systematic review. J Med Syst 2020 Apr 03;44(5):100. [doi: 10.1007/s10916-020-01556-z] [Medline: 32246206]
7. Ward MM, Jaana M, Natafgi N. Systematic review of telemedicine applications in emergency rooms. Int J Med Inform 2015 Sep;84(9):601-616. [doi: 10.1016/j.ijmedinf.2015.05.009] [Medline: 26072326]
8. Kvedar J, Coye MJ, Everett W. Connected health: a review of technologies and strategies to improve patient care with telemedicine and telehealth. Health Aff (Millwood) 2014 Feb;33(2):194-199. [doi: 10.1377/hlthaff.2013.0992] [Medline: 24493760]
9. Mitrasinovic S, Camacho E, Trivedi N, Logan J, Campbell C, Zilinyi R, et al. Clinical and surgical applications of smart glasses. Technol Health Care 2015;23(4):381-401. [doi: 10.3233/THC-150910] [Medline: 26409906]
10. Rogers H, Madathil KC, Agnisarman S, Narasimha S, Ashok A, Nair A, et al. A systematic review of the implementation challenges of telemedicine systems in ambulances. Telemed J E Health 2017 Sep;23(9):707-717. [doi: 10.1089/tmj.2016.0248] [Medline: 28294704]
11. Romare C, Skär L. Smart glasses for caring situations in complex care environments: scoping review. JMIR Mhealth Uhealth 2020 Apr 20;8(4):e16055 [FREE Full text] [doi: 10.2196/16055] [Medline: 32310144]
12. Klein GO, Singh K, von Heideken J. Smart glasses--a new tool in medicine. Stud Health Technol Inform 2015;216:901. [Medline: 26262203]
13. Dougherty B, Badawy SM. Using Google Glass in nonsurgical medical settings: systematic review. JMIR Mhealth Uhealth 2017 Oct 19;5(10):e159 [FREE Full text] [doi: 10.2196/mhealth.8671] [Medline: 29051136]
14. Wei NJ, Dougherty B, Myers A, Badawy SM. Using Google Glass in surgical settings: systematic review. JMIR Mhealth Uhealth 2018 Mar 06;6(3):e54 [FREE Full text] [doi: 10.2196/mhealth.9409] [Medline: 29510969]
15. Aldaz G, Shluzas LA, Pickham D, Eris O, Sadler J, Joshi S, et al. Hands-free image capture, data tagging and transfer using Google Glass: a pilot study for improved wound care management. PLoS One 2015 Apr 22;10(4):e0121179 [FREE Full text] [doi: 10.1371/journal.pone.0121179] [Medline: 25902061]
16. Klinker K, Wiesche M, Krcmar H. Development of a Smart Glass application for wound management. In: Proceedings of the 14th International Conference on Design Science Research in Information Systems and Technology: Extending the Boundaries of Design Science Theory and Practice. 2019 Presented at: DESRIST '19; June 4-6, 2019; Worcester, MA, USA p. 157-171. [doi: 10.1007/978-3-030-19504-5_11]
17. Apiratwarakul K, Cheung LW, Tiamkao S, Phungoen P, Tientanopajai K, Taweepworadej W, et al. Smart Glasses: a new tool for assessing the number of patients in mass-casualty incidents. Prehosp Disaster Med 2022 Aug;37(4):480-484 [FREE Full text] [doi: 10.1017/S1049023X22000929] [Medline: 35757837]
18. Cicero MX, Walsh B, Solad Y, Whitfill T, Paesano G, Kim K, et al. Do you see what I see? Insights from using google glass for disaster telemedicine triage. Prehosp Disaster Med 2015 Feb;30(1):4-8. [doi: 10.1017/S1049023X1400140X] [Medline: 25571779]

19. Broach J, Hart A, Griswold M, Lai J, Boyer EW, Skolnik AB, et al. Usability and reliability of smart glasses for secondary triage during mass casualty incidents. Proc Annu Hawaii Int Conf Syst Sci 2018 Jan 03;2018:1416-1422 [FREE Full text] [doi: 10.24251/hicss.2018.175] [Medline: 29398976]
20. Wrzesińska N. The use of smart glasses in healthcare - review. MEDtube Sci 2015 Dec;4(3):31-34 [FREE Full text]
21. Cakmakci O, Rolland J. Head-worn displays: a review. J Display Technol 2006 Sep;2(3):199-216. [doi: 10.1109/jdt.2006.879846]
22. Google Glass. Wikipedia - The Free Encyclopedia. URL: https://en.wikipedia.org/wiki/Google_Glass [accessed 2023-02-09]
23. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Int J Surg 2010;8(5):336-341 [FREE Full text] [doi: 10.1016/j.ijsu.2010.02.007] [Medline: 20171303]
24. Mankins JC. Technology readiness levels: a white paper. Office of Space Access and Technology, National Aeronautics and Space Administration. 1995 Apr 6. URL: https://aiaa.kavi.com/apps/group_public/download.php/2212/TRLs_MankinsPaper_1995.pdf [accessed 2023-02-09]
25. Engel DW, Dalton AC, Anderson K, Sivaramakrishnan C, Lansing C. Development of technology readiness level (TRL) metrics and risk measures. Pacific Northwest National Laboratory. Richland, WA, USA: U.S. Department of Energy; 2012 Oct. URL: https://www.pnnl.gov/main/publications/external/technical_reports/PNNL-21737.pdf [accessed 2023-02-09]
26. Polit DF, Beck CT. Nursing Research: Generating and Assessing Evidence for Nursing Practice. Philadelphia, PA, USA: Lippincott Williams & Wilkins; 2008.
27. Chai PR, Babu KM, Boyer EW. The feasibility and acceptability of Google Glass for teletoxicology consults. J Med Toxicol 2015 Sep;11(3):283-287 [FREE Full text] [doi: 10.1007/s13181-015-0495-7] [Medline: 26245879]
28. Noorian AR, Bahr Hosseini M, Avila G, Gerardi R, Andrle AF, Su M, et al. Use of wearable technology in remote evaluation of acute stroke patients: feasibility and reliability of a Google Glass-based device. J Stroke Cerebrovasc Dis 2019 Oct;28(10):104258. [doi: 10.1016/j.jstrokecerebrovasdis.2019.06.016] [Medline: 31296476]
29. Ponce BA, Menendez ME, Oladeji LO, Fryberger CT, Dantuluri PK. Emerging technology in surgical education: combining real-time augmented reality and wearable computing devices. Orthopedics 2014 Nov;37(11):751-757. [doi: 10.3928/01477447-20141023-05] [Medline: 25361359]
30. Gupta S, Boehme J, Manser K, Dewar J, Miller A, Siddiqui G, et al. Does wearable medical technology with video recording capability add value to on-call surgical evaluations? Surg Innov 2016 Oct;23(5):498-504. [doi: 10.1177/1553350616656278] [Medline: 27335083]
31. Demir F, Ahmad S, Calyam P, Jiang D, Huang R, Jahnke I. A next-generation augmented reality platform for Mass Casualty Incidents (MCI). J Usability Stud 2017 Aug;12(4):193-214 [FREE Full text]
32. Ho TC, Kolin T, Stewart C, Reid MW, Lee TC, Nallasamy S. Evaluation of high-definition video smart glasses for real-time telemedicine strabismus consultations. J AAPOS 2021 Apr;25(2):74.e1-74.e6. [doi: 10.1016/j.jaapos.2020.11.016] [Medline: 33901673]
33. Brewer ZE, Fann HC, Ogden WD, Burdon TA, Sheikh AY. Inheriting the learner's view: a Google Glass-based wearable computing platform for improving surgical trainee performance. J Surg Educ 2016;73(4):682-688. [doi: 10.1016/j.jsurg.2016.02.005] [Medline: 27137668]
34. Hashimoto DA, Phitayakorn R, Fernandez-del Castillo C, Meireles O. A blinded assessment of video quality in wearable technology for telementoring in open surgery: the Google Glass experience. Surg Endosc 2016 Jan;30(1):372-378. [doi: 10.1007/s00464-015-4178-x] [Medline: 25829065]
35. McCullough MC, Kulber L, Sammons P, Santos P, Kulber DA. Google Glass for remote surgical tele-proctoring in low- and middle-income countries: a feasibility study from Mozambique. Plast Reconstr Surg Glob Open 2018 Dec;6(12):e1999 [FREE Full text] [doi: 10.1097/GOX.0000000000001999] [Medline: 30656104]
36. Datta N, MacQueen IT, Schroeder AD, Wilson JJ, Espinoza JC, Wagner JP, et al. Wearable technology for global surgical teleproctoring. J Surg Educ 2015;72(6):1290-1295. [doi: 10.1016/j.jsurg.2015.07.004] [Medline: 26276303]
37. Martínez-Galdámez M, Fernández JG, Arteaga MS, Pérez-Sánchez L, Arenillas JF, Rodríguez-Arias C, et al. Smart glasses evaluation during the COVID-19 pandemic: first-use on neurointerventional procedures. Clin Neurol Neurosurg 2021 Apr 19;205:106655 [FREE Full text] [doi: 10.1016/j.clineuro.2021.106655] [Medline: 33962147]
38. Ye J, Zuo Y, Xie T, Wu M, Ni P, Kang Y, et al. A telemedicine wound care model using 4G with smart phones or smart glasses: a pilot study. Medicine (Baltimore) 2016 Aug;95(31):e4198 [FREE Full text] [doi: 10.1097/MD.0000000000004198] [Medline: 27495023]
39. Follmann A, Ohligs M, Hochhausen N, Beckers SK, Rossaint R, Czaplik M. Technical support by smart glasses during a mass casualty incident: a randomized controlled simulation trial on technically assisted triage and telemedical app use in disaster medicine. J Med Internet Res 2019 Jan 03;21(1):e11939 [FREE Full text] [doi: 10.2196/11939] [Medline: 30609988]
40. Drummond D, Arnaud C, Guedj R, Duguet A, de Suremain N, Petit A. Google Glass for residents dealing with pediatric cardiopulmonary arrest: a randomized, controlled, simulation-based study. Pediatr Crit Care Med 2017 Feb;18(2):120-127. [doi: 10.1097/PCC.0000000000000977] [Medline: 28165347]
41. Del Rio M, Meloni V, Frexia F, Cabras F, Tumbarello R, Montis S, et al. Augmented reality for supporting real time telementoring: an exploratory study applied to ultrasonography. In: Proceedings of the 2nd International Conference on

Medical and Health Informatics. 2018 Jun Presented at: ICMHI '18; June 8-10, 2018; Tsukuba, Japan p. 218-222. [doi: 10.1145/3239438.3239444]
42. Widmer A, Müller H. Using Google Glass to enhance pre-hospital care. Swiss Med Inform 2014 Sep 27;30:1-4. [doi: 10.4414/smi.30.00316]
43. Munusamy T, Karuppiah R, Bahuri NF, Sockalingam S, Cham CY, Waran V. Telemedicine via smart glasses in critical care of the neurosurgical patient-COVID-19 pandemic preparedness and response in neurosurgery. World Neurosurg 2021 Jan;145:e53-e60 [FREE Full text] [doi: 10.1016/j.wneu.2020.09.076] [Medline: 32956888]
44. Yoon H, Kim SK, Lee Y, Choi J. Google Glass-supported cooperative training for health professionals: a case study based on using remote desktop virtual support. J Multidiscip Healthc 2021 Jun 17;14:1451-1462 [FREE Full text] [doi: 10.2147/JMDH.S311766] [Medline: 34168458]
45. Diaka J, Van Damme W, Sere F, Benova L, van de Put W, Serneels S. Leveraging smart glasses for telemedicine to improve primary healthcare services and referrals in a remote rural district, Kingandu, DRC, 2019-2020. Glob Health Action 2021 Dec 06;14(1):2004729 [FREE Full text] [doi: 10.1080/16549716.2021.2004729] [Medline: 34889718]
46. Kujala S. User involvement: a review of the benefits and challenges. Behav Inf Technol 2003 Jan;22(1):1-16. [doi: 10.1080/01449290301782]
47. Klaassen B, van Beijnum BJ, Hermens HJ. Usability in telemedicine systems-a literature survey. Int J Med Inform 2016 Sep;93:57-69. [doi: 10.1016/j.ijmedinf.2016.06.004] [Medline: 27435948]
48. Kaipio J. Usability in healthcare: overcoming the mismatch between information systems and clinical work. Aalto University. 2011. URL: https://aaltodoc.aalto.fi/bitstream/handle/123456789/5041/isbn9789526043340.pdf?sequence=1 [accessed 2023-02-09]
49. Gosbee J, Klancher J, Arnecke B, Wurster H, Scanlon M. The role of usability testing in healthcare organizations. Proc Hum Factors Ergon Soc Annu Meet 2001 Oct;45(17):1308-1311. [doi: 10.1177/154193120104501711]
50. Schlosser P, Matthews B, Salisbury I, Sanderson P, Hayes S. Head-worn displays for emergency medical services staff: properties of prehospital work, use cases, and design considerations. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. 2021 May Presented at: CHI '21; May 8-13, 2021; Yokohama, Japan p. 40. [doi: 10.1145/3411764.3445614]
51. Sittig DF, Singh H. A new socio-technical model for studying health information technology in complex adaptive healthcare systems. In: Patel VL, Kannampallil TG, Kaufman DR, editors. Cognitive Informatics for Biomedicine: Human Computer Interaction in Healthcare. Cham, Switzerland: Springer; 2015:59-80.
52. Metzger J, Welebob E, Bates DW, Lipsitz S, Classen DC. Mixed results in the safety performance of computerized physician order entry. Health Aff (Millwood) 2010 Apr;29(4):655-663. [doi: 10.1377/hlthaff.2010.0160] [Medline: 20368595]
53. Harrington L, Kennerly D, Johnson C. Safety issues related to the electronic medical record (EMR): synthesis of the literature from the last decade, 2000-2009. J Healthc Manag 2011;56(1):31-44. [Medline: 21323026]
54. Magrabi F, Ong MS, Runciman W, Coiera E. Using FDA reports to inform a classification for health information technology safety problems. J Am Med Inform Assoc 2012;19(1):45-53 [FREE Full text] [doi: 10.1136/amiajnl-2011-000369] [Medline: 21903979]
55. Vogelsmeier AA, Halbesleben JR, Scott-Cawiezell JR. Technology implementation and workarounds in the nursing home. J Am Med Inform Assoc 2008;15(1):114-119 [FREE Full text] [doi: 10.1197/jamia.M2378] [Medline: 17947626]
56. Or C, Dohan M, Tan J. Understanding critical barriers to implementing a clinical information system in a nursing home through the lens of a socio-technical perspective. J Med Syst 2014 Sep;38(9):99. [doi: 10.1007/s10916-014-0099-9] [Medline: 25047519]
57. Sittig D, Kahol K, Singh H. Sociotechnical evaluation of the safety and effectiveness of point-of-care mobile computing devices: a case study conducted in India. In: Sittig DF, editor. Electronic Health Records: Challenges in Design and Implementation. Palm Bay, FL, USA: Apple Academic Press; 2013:115-133.
58. Zhang Z, Ramiya Ramesh Babu NA, Adelgais K, Ozkaynak M. Designing and implementing smart glass technology for emergency medical services: a sociotechnical perspective. JAMIA Open 2022 Dec;5(4):ooac113 [FREE Full text] [doi: 10.1093/jamiaopen/ooac113] [Medline: 36601367]
59. Zhang Z, Joy K, Harris R, Ozkaynak M, Adelgais K, Munjal K. Applications and user perceptions of smart glasses in emergency medical services: semistructured interview study. JMIR Hum Factors 2022 Feb 28;9(1):e30883 [FREE Full text] [doi: 10.2196/30883] [Medline: 35225816]
60. Lacy AM, Bravo R, Otero-Piñeiro AM, Pena R, De Lacy FB, Menchaca R, et al. 5G-assisted telementored surgery. Br J Surg 2019 Nov;106(12):1576-1579. [doi: 10.1002/bjs.11364] [Medline: 31483054]

Abbreviations

AR: augmented reality
EM: emergency medicine
FOV: field of view
HIPAA: Health Insurance Portability and Accountability Act

HIT: health IT
LMIC: low- and middle-income country
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
TRL: technology readiness level

Edited by C Lovis, G Eysenbach; submitted 08.11.22; peer-reviewed by Y Lee, M Eckert, MA Islam; comments to author 07.12.22; revised version received 15.01.23; accepted 31.01.23; published 28.02.23.

Please cite as:
Zhang Z, Bai E, Joy K, Ghelaa PN, Adelgais K, Ozkaynak M
Smart Glasses for Supporting Distributed Care Work: Systematic Review
JMIR Med Inform 2023;11:e44161
URL: https://medinform.jmir.org/2023/1/e44161
doi:10.2196/44161
PMID:36853760

Review

Mining Sensor Data to Assess Changes in Physical Activity Behaviors in Health Interventions: Systematic Review

Claudio Diaz[1], MBE; Corinne Caillaud[2], PhD; Kalina Yacef[1], PhD

[1]School of Computer Science, The University of Sydney, Sydney, Australia

[2]Charles Perkins Centre, School of Medical Sciences, The University of Sydney, Sydney, Australia

Corresponding Author:
Kalina Yacef, PhD
School of Computer Science
The University of Sydney
Building J12/1 Cleveland Street
Camperdown NSW
Sydney, 2006
Australia
Phone: 61 (02) 9351 2222
Email: kalina.yacef@sydney.edu.au

Abstract

Background: Sensors are increasingly used in health interventions to unobtrusively and continuously capture participants' physical activity in free-living conditions. The rich granularity of sensor data offers great potential for analyzing patterns and changes in physical activity behaviors. The use of specialized machine learning and data mining techniques to detect, extract, and analyze these patterns has increased, helping to better understand how participants' physical activity evolves.

Objective: The aim of this systematic review was to identify and present the various data mining techniques employed to analyze changes in physical activity behaviors from sensors-derived data in health education and health promotion intervention studies. We addressed two main research questions: (1) What are the current techniques used for mining physical activity sensor data to detect behavior changes in health education or health promotion contexts? (2) What are the challenges and opportunities in mining physical activity sensor data for detecting physical activity behavior changes?

Methods: The systematic review was performed in May 2021 using the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines. We queried the Association for Computing Machinery (ACM), IEEE Xplore, ProQuest, Scopus, Web of Science, Education Resources Information Center (ERIC), and Springer literature databases for peer-reviewed references related to wearable machine learning to detect physical activity changes in health education. A total of 4388 references were initially retrieved from the databases. After removing duplicates and screening titles and abstracts, 285 references were subjected to full-text review, resulting in 19 articles included for analysis.

Results: All studies used accelerometers, sometimes in combination with another sensor (37%). Data were collected over a period ranging from 4 days to 1 year (median 10 weeks) from a cohort size ranging between 10 and 11615 (median 74). Data preprocessing was mainly carried out using proprietary software, generally resulting in step counts and time spent in physical activity aggregated predominantly at the daily or minute level. The main features used as input for the data mining models were descriptive statistics of the preprocessed data. The most common data mining methods were classifiers, clusters, and decision-making algorithms, and these focused on personalization (58%) and analysis of physical activity behaviors (42%).

Conclusions: Mining sensor data offers great opportunities to analyze physical activity behavior changes, build models to better detect and interpret behavior changes, and allow for personalized feedback and support for participants, especially where larger sample sizes and longer recording times are available. Exploring different data aggregation levels can help detect subtle and sustained behavior changes. However, the literature suggests that there is still work remaining to improve the transparency, explicitness, and standardization of the data preprocessing and mining processes to establish best practices and make the detection methods easier to understand, scrutinize, and reproduce.

(JMIR Med Inform 2023;11:e41153) doi:10.2196/41153

KEYWORDS

activity tracker; wearable electronic devices; fitness trackers; data mining; artificial intelligence; health; education; behavior change; physical activity; wearable devices; trackers; health education; sensor data

Introduction

Wearable sensors are increasingly employed in health interventions because of their ability to track participants' physical activity (PA) in an unobtrusive, continuous, and precise manner under free-living conditions [1]. In the context of health promotion, sensor data are commonly used to objectively assess interventions by monitoring PA changes and progress toward compliance with public health PA guidelines [2].

The rich data captured by activity sensors contain information about the participants' PA, potentially unlocking valuable insights into PA behaviors and patterns [3]. These insights can help to advance the understanding of how interventions affect PA behaviors and how behaviors change, thereby scaffolding the design of future interventions, and enhancing their outcomes, efficacy, and adherence.

In the last decade, a growing number of artificial intelligence and data mining models and techniques have been developed to detect and extract these latent PA patterns beyond the typical summaries of pre- and postintervention daily steps or time spent in various PA levels. In this systematic review, we aimed to describe the data mining models and techniques currently used to detect PA with a focus on behavior changes. We discuss their value, identify gaps or challenges, and highlight opportunities. The following research questions (RQs) guided this review:

RQ1: What are the current techniques used for mining PA sensor data to detect behavior changes in health education or health promotion contexts?

RQ1.1 What are the types of sensors used and what data are collected?

RQ1.2 How are data preprocessed?

RQ1.3 What features are used to detect behavior changes?

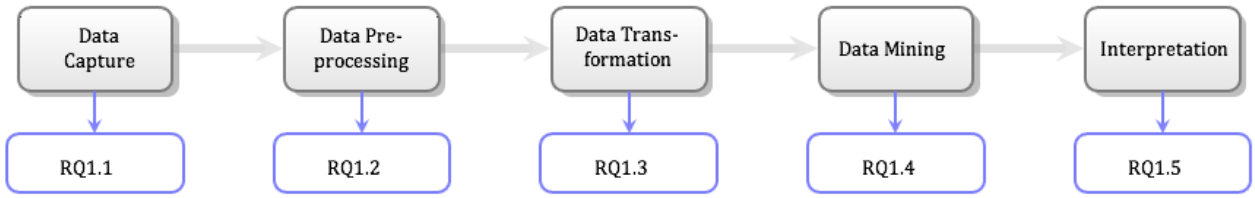
RQ1.4 What are the data mining models and techniques used to detect behavior changes?

RQ1.5 What are the interpretation of data mining models used for?

RQ2: What are the challenges and opportunities in mining PA sensor data for detecting PA behavior changes?

The RQ1 subquestions were established following the reasoning and order of the process of knowledge discovery in databases [4]. Figure 1 summarizes this process and maps each step with the relevant RQ1 subquestion.

Figure 1. Knowledge discovery in database steps (in grey) and research question 1 (RQ1) subquestions (in blue).



Methods

Design

For this systematic review, we followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [5] and used the Rayyan QCRI web application [6] to manage the review process. We identified studies by searching the Association for Computing Machinery (ACM), IEEE Xplore, ProQuest, Scopus, Web of Science, Education Resources Information Center (ERIC), and Springer digital libraries. We also searched Google Scholar to identify grey literature and extracted the first 100 results. For this scholarly reference search, we used the following query: (education OR promotion OR "behaviour change") AND ("data mining" OR "machine learning" OR "artificial intelligence") AND (sensor OR accelerometer OR tracker OR wearable) AND "physical activity" AND health. All extracted scholarly references had been added to the database at the latest on the search day (May 28, 2021). The inclusion and exclusion criteria are presented in Textbox 1.

Textbox 1. Inclusion and exclusion criteria for article selection in the review.

Inclusion criteria

- Full-length articles
- Peer-reviewed articles in journals or conference papers
- Articles that used data mining techniques for data from physical activity (PA) wearable sensors
- Articles that included PA data
- Articles on applied health education/promotion or on behavior change scenarios
- Articles that used well-known data mining techniques such as classification, regression, clustering, association, and sequence algorithms, as well as specific algorithms to model PA data

Exclusion criteria

- Use of analytics without data mining
- Studies on animals (eg, accelerometers on dogs)
- Self-quantification without a health education or health motivation component
- Dissertations and theses, due to lack of a peer review process
- Systematic reviews, reviews, and meta-analyses
- Health care applications without a health education or motivation for behavior change component
- Specific movement detection (abnormal gait, falls)
- Aid for sport training (eg, maintaining heart rate, postures, specific movements)

Search Outcome

The number of references extracted from each electronic database is summarized in Table 1.

Following the PRISMA methodology, we retrieved 4388 references from the sources listed in Table 1. We then removed 415 duplicates, leaving 3973 unique references that were screened by reading their titles and abstracts. Using the inclusion/exclusion criteria (Textbox 1), we excluded 3688 references and selected 285 publications. After full-text reading, we excluded 266 references: 33 on activity recognition, 5 on data mining, 24 on systems, 31 on rehabilitation, 39 not on behavior changes, 54 without data mining, 51 not on health education/promotion, 13 not on PA, and 16 reviews. At the end of the selection process (summarized in Figure 2), we retained 19 references for this systematic review.

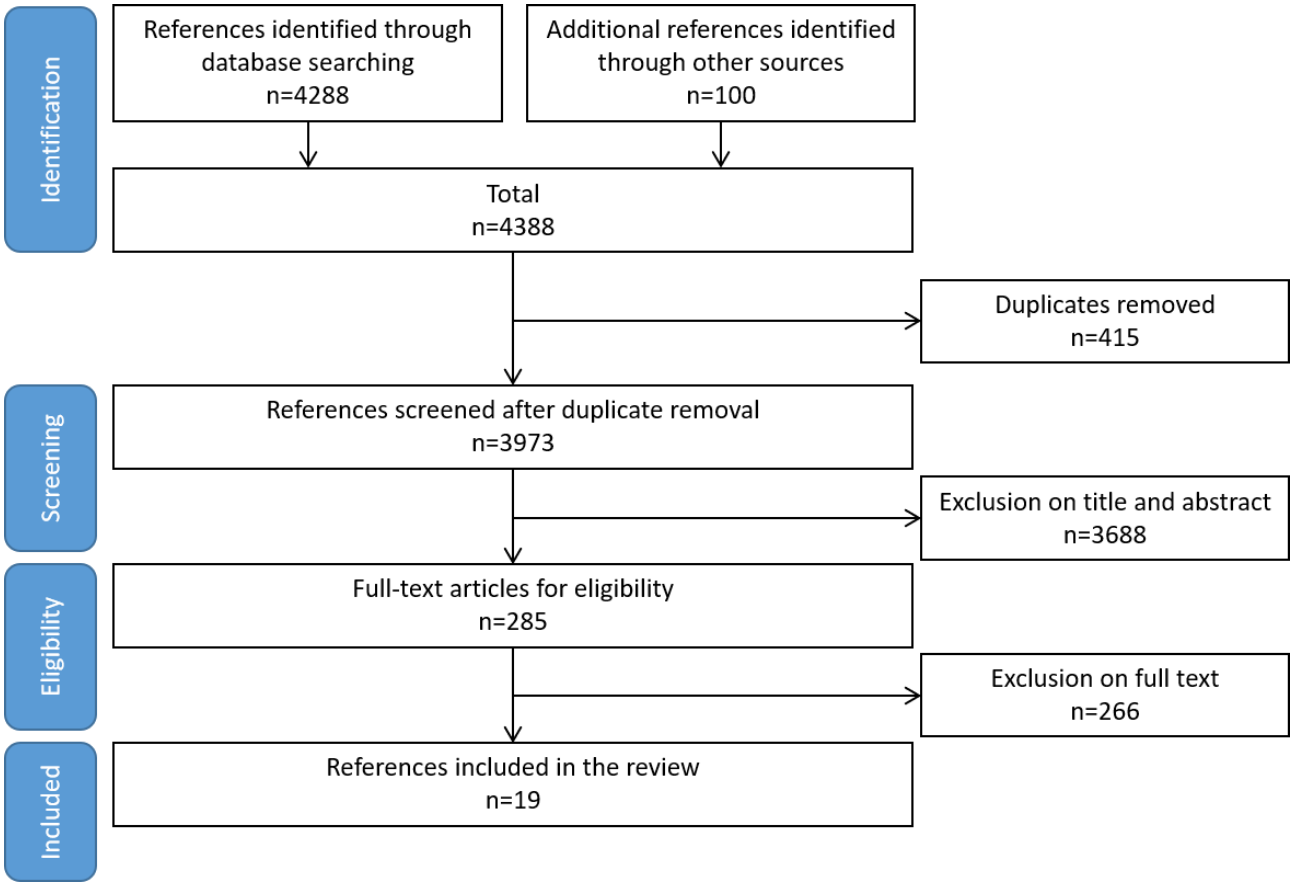
Table 1. Number of references extracted from each database.

Database	Query result, n
ACM[a]	584
IEEE Xplore	12
ProQuest	1678
Scopus	44
Web of Science	16
ERIC[b]	2
Springer	1952
Google Scholar	100

[a]ACM: Association for Computing Machinery.

[b]ERIC: Education Resources Information Center.

Figure 2. Study inclusion flowchart according to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) methodology.



Results

Overview

The 19 included articles were published between 2013 and 2021. Their number per year increased from 1 in 2013 to 2 in 2017 and up to 5 in 2018. Subsequently, the number of publications decreased to a mean of 3 per year.

The selected articles were published in conferences and journals focused on five different themes (Table 2): medical and public health, medical and health informatics, human-computer interactions, physical human behavior, and engineering and science. The three most popular themes were medical and health informatics, human-computer interactions, and engineering and science (15/19, 79%). Among the included articles, four were published in JMIR publications: three in JMIR mHealth and uHealth and one in JMIR Public Health and Surveillance.

Table 2. Conference proceedings and journals in which the included articles were published (N=19).

Conference or journal	Reference
Medical and public health	
BMJ Open	Aguilera et al [7]
Public Health Nutrition	Lee et al [8]
Medical and health informatics	
JMIR mHealth and uHealth	Zhou et al [9], Rabbi et al [10], Galy et al [11]
JMIR Public Health and Surveillance	Fukuoka et al [12]
Journal of Biomedical Informatics	Sprint et al [13]
Human-computer interactions	
Proceedings of the ACM on Human-Computer Interaction	Zhu et al [14]
User Modeling and User-Adapted Interaction	Gasparetti et al [15]
Journal of Ambient Intelligence and Humanized Computing	Batool et al [16]
Multimedia Tools and Applications	Angelides et al [17]
Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization	Schäfer et al [18]
Physical human behavior	
Journal of Behavioral Medicine	Forman et al [19]
Journal of Electromyography and Kinesiology	Hermens et al [20]
Engineering and science	
Applied Sciences	Chen et al [21]
Sensors	Dijkhuis et al [22]
Springer Proceedings in Complexity	Mollee et al [23]
IEEE Access	Diaz et al [24]
International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems	Mollee and Klein [25]

Sensor Types and Data Capture

The characteristics of the sensors (eg, number and type) used to capture PA behaviors and of the collected raw data are summarized in Table 3.

The length of data recordings varied between 4 days and 1 year, with a median of 70 days. Recording lasted ≤7 days in two studies, between 3 and 5 weeks in six studies, between 10 and 16 weeks in eight studies, and ≥6 months in three studies.

The number of participants varied between 10 and 11,615, with <30 in five studies, between 30 and 299 in 10 studies, and ≥300 participants in four studies.

All included studies used accelerometer sensors. We could categorize these devices into three groups: (1) commercial wrist-worn wearable accelerometers that are consumer-grade devices with a sample rate between 30 Hz and 60 Hz, such as Fitbits [13,14,19,22,25], Samsung Gear [17], and Nokia [15]; (2) smartphone accelerometers with a sample rate usually set to 50 Hz and up to 100 Hz, in which data were collected via an app installed in the smartphone [7,9,10,16,18]; and (3) scientifically validated wearable accelerometers with a sample rate up to 100 Hz, such as ActiGraph [8], GENEActiv [11,24], and other devices developed for health care [12,20].

In 7 out of the 19 (37%) selected studies, accelerometers were used with other sensors such as GPS tracking [10,16,17], compass position tracking [17,20], heart rate trackers [17,21], and smart scales [15,19].

The recorded raw data varied in function of the sensor characteristics, including sampling frequency, accuracy, and axis number. Moreover, other sensor features such as battery duration and storage capacity affected the recording length. For instance, a long battery life and high storage capacity enable longer recording without interruptions. Table 4 summarizes the number of participants and data recording duration for the included studies.

Table 3. Number of sensors, device type and model used, and raw data generated.

Sensor type	Device and model	Raw data	Reference
Accelerometer	ActiGraph GT1M uniaxial	Uniaxial accelerometry	Lee et al [8]
Accelerometer	GENEActiv triaxial accelerometer	Gravity-subtracted signal vector magnitudes (SVMgs) per second	Galy et al [11], Diaz et al [24]
Accelerometer	Generic device from the mobile phone	Acceleration (sample rate not specified)	Aguilera et al [7], Zhou et al [9]
Accelerometer	Triaxial accelerometer (HJA-350IT, Active Style Pro, Omron Healthcare Co, Ltd)	Triaxial acceleration (6 Hz)	Fukuoka et al [12]
Accelerometer	Fitbit	Triaxial acceleration (sample rate not specified)	Zhu et al [14]
Accelerometer	Fitbit Flex	Triaxial acceleration (sample rate not specified)	Dijkhuis et al [22]
Accelerometer	Fitbit Charge HR and Fitbit Flex	Triaxial acceleration (sample rate not specified)	Sprint et al [13]
Accelerometer	Fitbit One	Triaxial acceleration (sample rate not specified)	Mollee and Klein [25]
Accelerometer	Not specified	Not specified	Mollee et al [23]
Accelerometer	Smartphone and Actigraph (GT3X model)	Triaxial acceleration (sample rate not specified)	Schäfer et al [18]
Accelerometer and heart rate monitor	Mix of devices and models	Accelerometry, heart rate monitor, PA[a] information, and user information (sample rate not specified)	Chen et al [21]
Accelerometer, GPS, self-log PA, and food	Smartphone	Smartphone accelerometry, GPS data, PA and food logs with sample rate specified	Rabbi et al [10]
Activity tracker, smart scale, and smartphone (what they ate and drank in the Fitbit app)	Fitbit Flex 2 activity tracker, Yunmai smart scale, smartphone	Accelerometry, weight and food logs (sample rate not specified)	Forman et al [19]
Accelerometer, gyroscope, and magnetic compass	ProMove-3D (developed by Inertia Technology)	Accelerometry (sample rate not specified)	Hermens et al [20]
Accelerometer and GPS	Smartphone	Accelerometry and GPS (sample rate not specified)	Batool et al [16]
Triaxial accelerometer, heart rate monitor, GPS, 3-axis gyroscope, digital compass, altimeter, light sensor	Samsung Gear Fit and Fitbit Surge	Accelerometry, heart rate data, GPS, 3-axis gyroscopes, digital compass, altimeter, light sensor (sample rate not specified)	Angelides et al [17]
Accelerometer, heart rate, and smart scale	Nokia; models not specified	Accelerometer, heart rate data, and smart scale (sample rate not specified)	Gasparetti et al [15]

[a]PA: physical activity.

Table 4. Length of data recording and number of participants among the included studies.

Length of recording	Participants, n	Reference
1 to 7 days		
4 days	1714	Lee et al [8]
7 days	215 (women)	Fukuoka et al [12]
1 to 5 weeks		
3 weeks	17	Rabbi et al [10]
3 weeks	48	Zhu et al [14]
1 month	14	Angelides et al [17]
4 weeks	24 (adolescents)	Galy et al [11]
4 weeks	74 (children)	Schäfer et al [18]
5 weeks	87 (children)	Diaz et al [24]
6 to 20 weeks		
10 weeks	11	Schäfer et al [18]
10 weeks	64	Zhou et al [9]
3 months	10	Hermens et al [20]
12 weeks	48	Dijkhuis et al [22]
12 weeks	108	Mollee and Klein [25]
3 months	269	Chen et al [21]
12 weeks	2472	Mollee et al [23]
16 weeks	52	Forman et al [19]
21 weeks to 1 year		
6 months	276	Aguilera et al [7]
6 months	500	Batool et al [16]
1 year	11,615	Gasparetti et al [15]

Data Preprocessing

Raw data extracted from sensors need to be transformed into variables that will contribute to generating the input features for data mining models to detect PA behavior changes. Table 5 provides a summary of the initial transformation and the resulting preprocessed data.

The preprocessing of the raw data from sensors was carried out in two ways. The first approach was to use proprietary programs to transform the sensors' data directly into the resulting preprocessed data, without specifying whether there was an initial preprocessing stage such as that used to generate steps, metabolic equivalents (METs), calories, heart rate, or exercise characteristics (type, duration, distance, or frequency). The second approach was to produce intermediate data that were then transformed in the resulting preprocessed data using a custom preprocessing tool. For instance, to generate PA levels (PALs), raw data were first transformed into MET, activity classes, or signal vector magnitudes.

The resulting preprocessed data were mainly activity characteristics (step count, PAL, integrals of the moduli of acceleration, activity types, duration, distance travelled, and frequency) and energy expenditure (MET and calories). Step count from smartphones and commercial wrist-worn devices was the most frequent, followed by PAL from research-grade devices.

The resulting preprocessed data were aggregated at different time levels (Table 6). Day and minutes were the most frequent time levels of aggregation. Generally, PAL and MET were aggregated per minute. Calories and step counts were calculated per day.

Table 5. Summary of data preprocessing variables.

Resulting/initial preprocessing	Reference
Steps: unknown (proprietary program)	[7,9,13-15,17,19,22,25]
Metabolic equivalents: unknown (proprietary program)	[21]
Calories: unknown (proprietary program)	[10,17,19]
Exercise characteristics: unknown (proprietary program)[a]	[10,17,21]
Sleeping time: unknown (proprietary program)	[15,17]
Weight: unknown (proprietary program)	[15,19]
Heart rate: unknown (proprietary program)	[17,21]
Physical activity (PA) levels	
Signal vector magnitudes	[11,24]
PA counts	[8]
Metabolic equivalents	[12]
Activity classes	[18]
Not specified	[23]
Integrals of the moduli of acceleration signals	[20]
Actual activity level; not specified[b]	[16]

[a]Type, duration, distance, frequency.

[b]Definition of activity level was not specified.

Table 6. Aggregation level of the resulting preprocessed data.

Reference	Month	Week	Day	Hour	Minute	Seconds	Not specified
Angelides et al [17]	✓	✓	✓	✓			
Zhou et al [9]			✓				
Aguilera et al [7]			✓				
Zhu et al [14]			✓				
Mollee and Klein [25]			✓				
Forman et al [19]			✓				
Gasparetti et al [15]			✓				
Chen et al [21]	✓				✓		
Dijkhuis et al [22]				✓			
Lee et al [8]					✓		
Fukuoka et al [12]					✓		
Sprint et al [13]					✓		
Schäfer et al [18]					✓		
Diaz et al [24]						✓	
Galy et al [11]						✓	
Mollee et al [23]							✓
Hermens et al [20]							✓
Rabbi et al [10]							✓
Batool et al [16]							✓

Features Used to Detect and Extract Behavior Changes

The features of the data mining models were mostly generated from the sensors' preprocessed data and, in some cases, from other sources (nonsensor data). Table 7 provides the features categorized with respect to the function of their source: accelerometers, other sensors, and nonsensor devices.

Most of the included articles used descriptive statistics to present the preprocessed data as features, for instance total number of steps per day [9,11,14,17,25], mean number of steps per day [17], or PA count per hour [8]. Other studies created windows or segments of time to calculate PA characteristics, including segments of steps or sleep [15] and PA bouts [13,24]. Other articles used the preprocessed data to calculate the participants' step achievements such as whether they reached their step goal [9,11,19,23]. Zhu et al [13] used more complex features such as the ratio between the most active and least active period or the circadian rhythm strength.

In addition to the features derived from sensors, others were created from measurements carried out during the intervention by scientists, such as the number of days that a person participated in the intervention [19] and anthropometric [7,21] or psychological [14,16,25] characteristics. Data were collected through surveys/questionnaires or interviews with participants.

Table 7. Features used for data mining to detect behavior changes.

Reference	Features derived from accelerometers	Features derived from other sensors	Features derived from nonsensor devices
Aguilera et al [7]	Number of minutes of activity in the last day, cumulative number of minutes of activity this week, fraction of activity goal, fraction versus expected activity goal at this point in the week	Number of days since each feedback message was sent	Age, gender, language, 8-item Patient Health Questionnaire (depression) score
Hermens et al [20]	Not specified	Not specified	Not specified
Chen et al [21]	Monthly mean metabolic equivalent of task, effective exercise time, type, frequency	Monthly mean exercise and resting heart rate	Gender, height, weight, age
Forman et al [19]	Days where PA[a] goal is met	Sum of days with self-monitored weight, days with self-monitored eating, days where calorie goal is met, weight loss in pounds	Number of days in the intervention period
Gasparetti et al [15]	Consecutive daily segments of steps, consecutive daily segments of sleep	__[b]	—
Batool et al [16]	Actual activity level	—	Desired activity level, intention (attitude, subjective norms, perceived behavioral control), habit, and 16 demographic features (eg, age, gender, marital status)
Dijkhuis et al [22]	Hour of the workday, number of steps for that hour, number of steps in the past hour, total number of steps up to that hour, mean number of steps of workdays	—	—
Rabbi et al [10]	PA frequency and calories	—	—
Zhou et al [9]	Daily steps and goal	—	—
Angelides et al [17]	Total and mean hourly, daily, weekly, and monthly sleep duration; sleep calories; exercise duration; exercise distance; exercise calories; step count; step distance; step calories; BMI; and basal metabolic rate	—	Height (cm), weight (kg), age, gender
Diaz et al [24]	Hourly and daily frequency, and mean time spent in moderate to vigorous PA bouts of at least 3, 10, and 30 seconds, and in sedentary bouts of at least 60, 120, and 300 seconds	—	—
Galy et al [11]	Total daily time spent in light/moderate/vigorous PA, total daily number of steps, and a binary goal achievement feature	—	—
Fukuoka et al [12]	Mean metabolic equivalent of tasks per minute, mean moderate-to-vigorous PA per minute	—	—
Lee et al [8]	24-hour mean PA count on weekdays and 24-hour mean PA count on weekends	—	—
Sprint et al [13]	Steps, PAL[c] and bouts count, mean, percentages, ratios and SD. Circadian rhythm time-series statistics and texture features from an image-processing technique	—	—
Mollee et al [23]	Impact of online community (sharing my PAL with peers), target PAL and goal achievement	—	—
Schäfer [18]	PAL per minute	—	—

Reference	Features derived from accelerometers	Features derived from other sensors	Features derived from nonsensor devices
Mollee and Klein [25]	Daily steps	—	Psychological questionnaire scores for self-efficacy, barriers, social norm, long-term goals, intentions, satisfaction, outcome expectations
Zhu et al [14]	Daily steps	Motivation to exercise (Likert scale)	Iowa-Netherlands Comparison Orientation Measure-23 (INCOM-23) for social comparison (psychometrics)

[a]PA: physical activity.

[b]Not applicable.

[c]PAL: physical activity level.

Data Mining

Algorithm Overview

Table 8 summarizes the data mining methods and specific algorithms used in the selected articles.

Clustering was the most used method, particularly the K-means algorithm. Indeed, in health interventions, the PA performed by each participant varies in duration, form, and intensity. Therefore, an algorithm that clusters PA behaviors is required to analyze them. The unsupervised K-means algorithm is suitable for this task. Indeed, due to its simplicity and ease of use, this is one of the most popular options for data mining [26]. Decision-making algorithms and classifiers were the second most used methods. Both rely on supervised algorithms that use PA characteristics as a method for predicting when and/or what information must be delivered to individual participants for increasing their PA. Other algorithms were also tested to extract PA behaviors, such as social cognitive and contagion models, PA windows permutations, and recommendation algorithms.

Table 8. Data mining methods and algorithms.

Data mining method and algorithm	Reference
Classifiers	
K-nearest neighbor and support vector machine	[20]
Random forest	[22]
Random forest and weighted score	[18]
Shallow neural networks	[16]
Clustering techniques	
K-means	[8,11,12,24]
Agglomerative	[21]
Partitioning around medoids and reinforcement learning	[15]
Decision-making algorithms	
Multiarmed bandit	[10]
Multiarmed bandit upper confidence bound	[19]
Reinforcement learning multiarmed bandit	[7,9]
Behavioral analytics algorithm	
MAB[a]	[14]
Social cognitive model for predicting exercise behavior change	[25]
Social contagion model combined with a linear model	[23]
Physical activity change detection: small window permutation-based change detection in activity routine	[13]
Recommendation: genetic algorithms and Pareto optimality	[17]

[a]MAB: multiarmed bandit.

Classifiers

Hermens et al [20] used a k-nearest neighbor model and a support vector machine to determine whether a specific time of the day was suitable for sending a motivational message to optimize adherence to the intervention. Dijkhuis et al [22] used a tree and tree-based ensemble algorithm classifiers to predict whether users will achieve their daily PA goal. On the basis of this prediction, a personalized PA coaching program was proposed. Forman et al [18] developed gamified personalized

feedback using a score model depending on the PA change detected from accelerometer data. Batool et al [16] predicted the likelihood that the PA level of a given patient was too low. They also predicted which patients were at higher risk of not adhering to the prescribed therapy to optimize their PA.

Clustering Techniques

Lee et al [8] grouped participants in two clusters on the basis of their step counts (one more active than the other), and analyzed them to better understand these PA patterns. Diaz et al [24] used a clustering-based approach for a more insightful analysis of the participants' PA behavior and of the nature of the PA behavior changes, if present. Galy et al [11] clustered PA levels and daily step goal achievement to assess the adherence to a health program. Fukuoka et al [12] identified PA clusters to analyze and compare sociodemographic features and cardiometabolic risks among participants belonging to these clusters. Chen et al [21] clustered the participants' PA, and then established a system to adapt the exercise program for the next week as a function of the individual PA behavior change. Gasparetti et al [15] clustered the participants' PA to generate groups of habits recommended by a system to the participants with the objective of changing their PA to obtain weight loss effects.

Decision-Making Algorithms

Rabbi et al [10] generated personalized suggestions in which users were asked to continue, avoid, or make small changes to their existing PA behaviors in order to help them reach their PA goals. Forman et al [19] developed an algorithm that could personalize and optimize the PAL during the intervention as a function of the amount of PA performed. Aguilera et al [7] generated personalized messages for participants in the intervention to increase their PA and consequently the intervention effectiveness. Zhou et al [9] adapted the step goal settings of the intervention depending on the PA behavior change. Zhu et al [14] personalized social comparison among participants to motivate them toward improving their PA behavior.

Social Cognitive Model

Mollee and Klein [25] developed a model that simulates changes in PALs over 2 to 12 weeks to optimize the participants' health outcome.

Social Contagion Model

Mollee et al [23] used a social contagion model to explain the PAL dynamics in a community.

PA Windows Permutations

Sprint et al [13] proposed a window-based algorithm to detect changes in segments of users' PA behavior to motivate progress toward their goals.

Recommendation Algorithms

Angelides et al [17] used genetic algorithms and Pareto optimality to compare the participants' and peer community's data to help participants interpret the PA data and to generate personal lifestyle improvement recommendations.

Interpretation of the Data Mining Models

Overview of Models

The resulting data mining models detecting PA behavior changes were used for several purposes, as summarized in Table 9 and below.

Table 9. Main uses of the resulting data mining models.

Main use	Reference
Personalized feedback	[7,10,15,16,18,20]
Personalized program	[9,19,21,22]
Support for self-reflection	[17]
Cohort analysis of the intervention impact on PA[a]	[8,11-13,24]
Analysis of the social component effects on PA	[14,23,25]

[a]PA: physical activity.

Personalized Feedback

The PA behavior changes extracted from participants' data were used to promote PA by creating and sending personalized messages that reported the behaviors and gave suggestions for achieving the previously established PA goals. For instance, Aguilera et al [7] built a system that detects the participants' PA behavior changes and generates personalized daily text messages with custom timing, frequency, and feedback about their step count/goal and motivational content. Hermens et al [20] built a system that chooses the best suitable time to send a message with personalized intention, content, and representation. Schäfer et al [18] created an app with gamified feedback where different avatars are awarded based on the participant's daily PA behavior. Gasparetti et al [15] suggested personalized PA patterns based on the participants' PA patterns. Batool et al [16] detected the participants' PA behavior while commuting and suggested how to increase it. Rabbi et al [10] generated personalized simple PA suggestions (continue, avoid, or make small changes).

Personalized Programs

The PA intervention program and objectives are adapted to each participant's needs. For instance, Chen et al [21] created a guided exercise prescription system that adapts as the participants' PA behavior changes. Similarly, Forman et al [19] changed the participant's exercise intensity suggestion depending on their PA behavior achievements. On the basis of

each participant's step count progress, Dijkhuis et al [22] suggested new daily step objectives. Zhou et al [9] used push notifications to deliver daily step goals.

Support for Self-Reflection

Algorithms can help participants to interpret their PA behavior changes. For example, Angelides et al [17] used an algorithm to assist in the interpretation of the participant's PA data by comparing them with those of the peer community and to generate personalized recommendations to achieve their daily goals.

Cohort Analysis of the Intervention Impact on PA

These algorithms detect PA behavior changes in participants that allow analyzing the intervention impact. For example, Fukuoka et al [12] determined PA patterns in women throughout the day that could help to develop more personalized interventions and guidelines. Diaz et al [24] analyzed the changes in PA behavior (bouts and frequency) during an intervention. Galy et al [11] tracked the participants' adherence to the international recommendations during an intervention. Lee et al [8] identified PA patterns associated with specific subgroups of people who participated in an intervention. Sprint et al [13] analyzed the participants' PA changes during an intervention by comparing multiple time windows.

Analysis of the Social Component Effect on PA

These algorithms analyze the psychosocial influences on the participants' PA. For example, Mollee et al [23] analyzed the PA dynamics in a community using a social contagion model. Mollee and Klein [25] analyzed the PA dynamics in a networked community using social cognitive theories, and Zhu et al [14] personalized social comparison during an intervention to increase the participants' PA.

The main uses can be classified in two groups. The first group, composed of 11 out of the 19 (58%) selected studies, aimed to generate personalized feedback/PA programs to scaffold and support PA behavior changes among participants. Indeed, researchers seem inclined to generate greater personalization because it increases the intervention efficiency, effectiveness, enjoyment, and reliability [27]. The second group, composed of 8 out of the 19 (42%) selected studies, sought to analyze the impact of interventions on the participants' PA. Specifically, these studies analyzed the intervention impact on PA at the cohort level to assess health education interventions, and analyzed participants' PA to show them their behaviors and help to understand them. The main objective of both groups was to explore how PA behavior patterns relate to the intervention effectiveness, which can add new evidence on how to create more effective interventions [28].

Discussion

Principal Findings

Summary

We found 19 articles about data mining models and techniques to detect PA behavior changes in health education or promotion studies, and their number has progressively increased over time. We here discuss the principal findings, identify opportunities and challenges for future research directions, and present the limitations of this systematic review. The Discussion is structured according to the RQs as a guide.

Opportunities and Challenges

Sensor Types and Data Capture

All selected studies used accelerometer sensors to capture PA behaviors. While 7 out of the 19 (37%) studies utilized accelerometers exclusively, the rest employed them with other sensors. Nonaccelerometer sensors capture additional information that may be relevant to PA (such as work/school schedule, itineraries, and sleep patterns [29]) and could yield auxiliary features for the data mining models. For instance, GPS sensors provide the number of kilometers and location of PA performed.

The median number of participants in the selected studies was 74, and participants were mainly young or middle-aged adults. This low number of participants and the skew toward adults may have generated biased data mining models that can detect and find behavior changes only in a specific population. Different population groups behave differently and should be studied independently. For instance, PA behaviors are different in children and adults [2]. Some of the studies focused on groups with specific PA behaviors, such as children [18,24], adolescents [11], and women [12]. However, some population groups with distinctive PA patterns, such as pregnant women [30] and people with health conditions or disabilities [31], may need custom detection models.

In 15 out of the 19 (79%) included studies, data were recorded for less than 3 months. Therefore, the current methods for detecting PA behavior changes have been developed mostly for capturing short-term patterns, making the conclusions valid only for short periods. To detect medium- and long-term PA behavior changes, studies with more extended recording periods are needed, such as the study by Gasparetti et al [15] based on data collected during 1 year. Moreover, new methods to detect extended (eg, annual or seasonal) PA patterns are required to study how the participants' behavior and habits change over time. An increase in the participants' number and recording length will lead to new challenges related to big data analysis, such as efficient data management and data mining processing speeds.

Data Preprocessing

Many of the selected studies used commercial accelerometers that allow only the retrieval of aggregated preprocessed data using proprietary software (ie, number of steps per minute), without being transparent on how data were preprocessed (ie, how steps were calculated from the accelerometry data). This data preprocessing black box makes it impossible to determine the quality of the captured PA data and makes the data mining results scientifically irreproducible. Conversely, in studies that used medical-grade accelerometers, the accelerometry data were explained in detail and the preprocessing steps were documented and referenced.

We found a lack of standard procedures for data preprocessing that made it challenging to compare the study results and conclusions. Indeed, if data are not preprocessed correctly, this could cause the transfer of incorrect information to the features and then to the data mining models. This could lead to the creation of inaccurate models, thus limiting the study validity. Data cleaning is a good example of this issue. Indeed, the best procedure to eliminate the nonwearing time remains unclear along with the impact on the accuracy of the resulting models. If nonwearing time is poorly removed, features can generate a PA underestimation by recognizing nonwearing time as sedentary behavior when it is not. Moreover, if sensor data concerning changes in accelerations while commuting by car or bus are not completely removed, they will be erroneously classified as steps, thereby overestimating PA in the model and in the conclusions. Similarly, sedentary activities could be overestimated if sleep time is not correctly removed.

Most of the selected studies aggregated information by day or minute. Although data aggregation is useful when comparing general features of PA behaviors, such as daily steps, this procedure may overlook subtle behavioral changes that can be crucial for detecting major PA behavior changes. For instance, if a person who walks every morning decides to change their behavior and starts to walk at night, the sum of daily steps will be the same, but this new behavior will not be detected. Conversely, it could be detected if the aggregation level is changed to the hour. To detect these and other subtle behavior changes, PA should be analyzed simultaneously at different aggregation levels, and new time frames should be created to match daily habits and behaviors, such as periods of the day (eg, morning, afternoon) or participants' office hours.

Features Used to Detect and Extract Behavior Changes

Most of the preprocessed data were transformed into features that are simple descriptive statistics, such as the total time spent at a specific PAL or the mean number of steps. These features are valuable to detect behavior changes, but they mainly capture the PA intensity and the PA presence or absence. Yet, PA has more valuable characteristics that vary during PA behavior changes and that can help to detect such behavior changes, such as the length of PAL bouts or the amount of time spent doing PA. These PA characteristics can be extracted from current sensor data. For instance, Galy et al [11] explored different moderate-to-vigorous PA bout lengths and Sprint et al [13] assessed the circadian rhythm. International PA guidelines can serve as inspiration to identify new PA features. For instance, according to World Health Organization recommendations, adults should perform muscle-strengthening activities (involving all major muscle groups) at moderate or higher intensity at least twice per week [2]. This calls for the creation of features that capture the muscle activity type, intensity, and frequency. Moreover, most of the included studies used only PA-derived features to detect behavior changes, and did not consider relevant non-PA data associated with PA changes, such as the participants' weight and quality of sleep. Some studies captured non-PA data, but they did not use them to detect PA changes. For instance, Rabbi et al [10] used only PA-derived data (PA frequency and calories burned) to detect behavior changes, although they also recorded the participants' food intake, thus excluding their caloric intake that is closely related to weight and the amount of PA participants are likely perform.

The use of simple descriptive statistics as features and the exclusion of non-PA data associated with behavior changes indicate that sensor data were underexploited and that the features used to detect PA behavior changes are still underdeveloped. Including new PA characteristics and new non-PA features could help to better understand the nature of PA changes and how these features influence PA behavior changes, ultimately increasing the model detection accuracy.

Data Mining Methods and Techniques

Most studies used off-the-shelf classifiers, clusters, and decision-making algorithms to detect PA behavior changes. We expected to find tailor-made algorithms because in health education settings, it is important to find specific PA patterns in participants of different classes who follow learning modules with different contents and with different PA goals. Moreover, we noticed that most authors did not explain how they chose the algorithms and did not specify the efficiency and accuracy of the models used for detecting PA behavior changes, raising uncertainty about how good they are at this task. This suggests that more efficient and accurate algorithms could be created and calls for more transparency in the algorithm choice process. Therefore, authors should explicitly describe the steps and methodology of new algorithms, and share their source codes to be scrutinized and to compare their detection accuracy. The creation of open accelerometry databases is also needed to enable benchmarking.

Interpretation of the Resulting Data Mining Models

The main uses of the data mining models focused on personalization, support for self-reflection, and analysis of PA behaviors. Model interpretation focused on generating personalization and support for promoting behavior changes. Personalized feedback and intervention programs were based mostly on the participants' PA data. The inclusion of additional information that may influence behavior changes (eg, contexts, schedules, social constraints, motivation, and weather) would allow for better interpretation and use of the detected behavior changes. Systems could exploit these additional data to improve the feedback delivery time and content, with positive effects on the effectiveness of health education programs and interventions. For instance, with the current models, a participant could receive an automatized personalized behavior change message that suggests taking a short walk, although it is snowing outside. This would decrease the likelihood of following the suggestion. However, if the system could be aware of the weather, the participant would receive this suggestion only after the weather conditions have improved, or a different suggestion that is more likely to trigger a behavior change at that point in time. Moreover, as the models relied mainly on PA features to model and interpret the behavior changes, only the physical dimension of the learning process in health education was incorporated in the models and their interpretation, leaving aside the knowledge dimension of the learning process. Learning management systems and intelligent tutoring systems already capture the knowledge dimension. Their integration would help to understand, in a comprehensive way, how participants learn,

and would enable the real-time monitoring of how PA behavior changes align with the intervention purpose. This would allow adapting each participant's content and learning objectives in real time, thereby improving instructions and learning, ultimately increasing the program or intervention effectiveness.

Most of the included studies generated complex output models that require detailed knowledge of how they were created to interpret the resulting patterns, making them difficult to understand for health scientists and any other scientist not familiar with machine learning. This is a common problem in interdisciplinary teams; however, an effort can be made to create more readable, intuitive, and easy-to-understand algorithms and methods, a goal that exists in related machine learning areas such as explainable artificial intelligence [32].

Limitations

Studies on wearable machine learning devices to detect changes in PA in health education have only started to be published in the last decade. As research is advancing, keywords are changing and new terms are created. Although we used a wide range of keywords in our query to include sensors, PA, and health education, we may have left some keywords out, and thus we may have missed some references. This may have also affected the initial reference screening process by title and abstract. We minimized this issue by testing several queries before starting our systematic review until we found the one we ultimately used. Another possible limitation in our search is that we might have omitted references listed only in other peer-reviewed databases (we searched only the most popular databases in engineering and computer science), such as medical databases (ie, PubMed). We mitigated this risk by including grey literature in our systematic review (see the Methods section).

Regarding the research subquestions and the review structure, we created research subquestions in line with the usual data mining process steps, but we certainly left some topics unaddressed. For instance, we did not address ethics, privacy, and security issues, or how data are filtered during preprocessing (eg, sleeping time or sensor nonuse). Although these are common substeps during the data mining process and including them would have made this systematic review more comprehensive, we preferred to limit this review only to the critical steps.

Conclusions

In the last 10 years, different methods have been developed to detect behavior changes in health education or health promotion contexts. These methods have been tested in small populations, are based on short data-recording periods, and rely mainly on accelerometry data. Incorporating information that is complementary to the participants' PA data would allow for creating more precise detection models, better interpreting these models, and understanding how participants learn and what triggers new behaviors. Exploring other data aggregation levels, in addition to days and minutes, could help to detect more subtle and long-term behavior changes. Fully describing the data preprocessing methods and the efficiency and accuracy of the behavior change detection models would help to better understand, scrutinize, and compare studies. Detection models were mainly used to generate personalized feedback and to provide support for promoting or maintaining behavior changes, but did not integrate the knowledge dimension of the learning process. Adding the knowledge dimension and creating easier-to-understand models could facilitate the interpretation of participants' behavior changes in a more comprehensive way, opening the way toward better and deeper analyses and personalization.

Acknowledgments

CD acknowledges the support of the Universidad Adolfo Ibáñez and Comisión Nacional de Investigación Científica y Tecnológica/Chilean National Commission for Scientific and Technological Research (CONICYT) "Becas Chile" Doctoral Fellowship program (grant number 72200111).

Conflicts of Interest

None declared.

References

1. Plasqui G, Bonomi AG, Westerterp KR. Daily physical activity assessment with accelerometers: new insights and validation studies. Obes Rev 2013 Jun 07;14(6):451-462. [doi: 10.1111/obr.12021] [Medline: 23398786]
2. Bull FC, Al-Ansari SS, Biddle S, Borodulin K, Buman MP, Cardon G, et al. World Health Organization 2020 guidelines on physical activity and sedentary behaviour. Br J Sports Med 2020 Dec 25;54(24):1451-1462 [FREE Full text] [doi: 10.1136/bjsports-2020-102955] [Medline: 33239350]
3. Rowlands AV. Moving forward with accelerometer-assessed physical activity: two strategies to ensure meaningful, interpretable, and comparable measures. Pediatr Exerc Sci 2018 Nov 01;30(4):450-456. [doi: 10.1123/pes.2018-0201] [Medline: 30304982]
4. Fayyad U, Piatetsky-Shapiro G, Smyth P. The KDD process for extracting useful knowledge from volumes of data. Commun ACM 1996 Nov;39(11):27-34. [doi: 10.1145/240455.240464]
5. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JP, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. J Clin Epidemiol 2009 Oct;62(10):1-34 [FREE Full text] [doi: 10.1016/j.jclinepi.2009.06.006] [Medline: 19631507]

6. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev 2016 Dec 05;5(1):210 [FREE Full text] [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
7. Aguilera A, Figueroa CA, Hernandez-Ramos R, Sarkar U, Cemballi A, Gomez-Pathak L, et al. mHealth app using machine learning to increase physical activity in diabetes and depression: clinical trial protocol for the DIAMANTE Study. BMJ Open 2020 Aug 20;10(8):e034723 [FREE Full text] [doi: 10.1136/bmjopen-2019-034723] [Medline: 32819981]
8. Lee PH, Yu Y, McDowell I, Leung GM, Lam T. A cluster analysis of patterns of objectively measured physical activity in Hong Kong. Public Health Nutr 2013 Aug;16(8):1436-1444. [doi: 10.1017/S1368980012003631] [Medline: 22894896]
9. Zhou M, Fukuoka Y, Mintz Y, Goldberg K, Kaminsky P, Flowers E, et al. Evaluating machine learning-based automated personalized daily step goals delivered through a mobile phone app: randomized controlled trial. JMIR Mhealth Uhealth 2018 Jan 25;6(1):e28 [FREE Full text] [doi: 10.2196/mhealth.9117] [Medline: 29371177]
10. Rabbi M, Pfammatter A, Zhang M, Spring B, Choudhury T. Automated personalized feedback for physical activity and dietary behavior change with mobile phones: a randomized controlled trial on adults. JMIR Mhealth Uhealth 2015 May 14;3(2):e42 [FREE Full text] [doi: 10.2196/mhealth.4160] [Medline: 25977197]
11. Galy O, Yacef K, Caillaud C. Improving Pacific adolescents' physical activity toward international recommendations: exploratory study of a digital education app coupled with activity trackers. JMIR Mhealth Uhealth 2019 Dec 11;7(12):e14854 [FREE Full text] [doi: 10.2196/14854] [Medline: 31825319]
12. Fukuoka Y, Zhou M, Vittinghoff E, Haskell W, Goldberg K, Aswani A. Objectively measured baseline physical activity patterns in women in the mPED Trial: cluster analysis. JMIR Public Health Surveill 2018 Feb 01;4(1):e10 [FREE Full text] [doi: 10.2196/publichealth.9138] [Medline: 29391341]
13. Sprint G, Cook DJ, Schmitter-Edgecombe M. Unsupervised detection and analysis of changes in everyday physical activity data. J Biomed Inform 2016 Oct;63:54-65 [FREE Full text] [doi: 10.1016/j.jbi.2016.07.020] [Medline: 27471222]
14. Zhu J, Dallal DH, Gray RC, Villareale J, Ontañón S, Forman EM, et al. Personalization Paradox in Behavior Change Apps. Proc ACM Hum Comput Interact 2021 Apr 22;5(CSCW1):1-21. [doi: 10.1145/3449190]
15. Gasparetti F, Aiello LM, Quercia D. Personalized weight loss strategies by mining activity tracker data. User Model User-Adap Inter 2019 Jul 26;30(3):447-476. [doi: 10.1007/s11257-019-09242-7]
16. Batool T, Vanrompay Y, Neven A, Janssens D, Wets G. CTASS: an intelligent framework for personalized travel behaviour advice to cardiac patients. J Ambient Intell Human Comput 2018 May 21;10(12):4693-4705. [doi: 10.1007/s12652-018-0847-7]
17. Angelides MC, Wilson LAC, Echeverría PLB. Wearable data analysis, visualisation and recommendations on the go using android middleware. Multimed Tools Appl 2018 Jun 4;77(20):26397-26448. [doi: 10.1007/s11042-018-5867-y]
18. Schäfer H, Bachner J, Pretscher S, Groh G, Demetriou Y. Study on motivating physical activity in children with personalized gamified feedback. 2018 Presented at: UMAP 18 26th Conference on User Modeling, Adaptation and Personalization; July 8-11, 2018; Singapore. [doi: 10.1145/3213586.3225227]
19. Forman EM, Kerrigan SG, Butryn ML, Juarascio AS, Manasse SM, Ontañón S, et al. Can the artificial intelligence technique of reinforcement learning use continuously-monitored digital data to optimize treatment for weight loss? J Behav Med 2019 Apr 25;42(2):276-290 [FREE Full text] [doi: 10.1007/s10865-018-9964-1] [Medline: 30145623]
20. Hermens H, op den Akker H, Tabak M, Wijsman J, Vollenbroek M. Personalized Coaching Systems to support healthy behavior in people with chronic conditions. J Electromyogr Kinesiol 2014 Dec;24(6):815-826. [doi: 10.1016/j.jelekin.2014.10.003] [Medline: 25455254]
21. Chen H, Chen F, Lin S. An AI-based exercise prescription recommendation system. Appl Sci 2021 Mar 16;11(6):2661. [doi: 10.3390/app11062661]
22. Dijkhuis TB, Blaauw FJ, van Ittersum MW, Velthuijsen H, Aiello M. Personalized physical activity coaching: a machine learning approach. Sensors 2018 Feb 19;18(2):623 [FREE Full text] [doi: 10.3390/s18020623] [Medline: 29463052]
23. Mollee JS, Araújo EFM, Manzoor A, van Halteren AT, Klein MCA. Explaining changes in physical activity through a computational model of social contagion. In: Gonçalves B, Menezes R, Sinatra R, Zlatic V, editors. Complex Networks VIII. CompleNet 2017. Springer Proceedings in Complexity. Cham: Springer; 2017:213-223.
24. Diaz C, Galy O, Caillaud C, Yacef K. A clustering approach for modeling and analyzing changes in physical activity behaviors from accelerometers. IEEE Access 2020;8:224123-224134. [doi: 10.1109/access.2020.3044295]
25. Mollee JS, Klein MCA. Empirical validation of a computational model of influences on physical activity behavior. In: Benferhat S, Tabia K, Ali M, editors. Advances in artificial intelligence: from theory to practice. IEA/AIE 2017. Lecture Notes in Computer Science, vol 10351. Cham: Springer; 2017:353-363.
26. Wu X, Kumar V, Ross Quinlan J, Ghosh J, Yang Q, Motoda H, et al. Top 10 algorithms in data mining. Knowl Inf Syst 2007 Dec 4;14(1):1-37. [doi: 10.1007/s10115-007-0114-2]
27. Cheung KL, Durusu D, Sui X, de Vries H. How recommender systems could support and enhance computer-tailored digital health programs: a scoping review. Digit Health 2019 Jan 24;5:2055207618824727 [FREE Full text] [doi: 10.1177/2055207618824727] [Medline: 30800414]
28. Williams SL, French DP. What are the most effective intervention techniques for changing physical activity self-efficacy and physical activity behaviour--and are they the same? Health Educ Res 2011 Apr 14;26(2):308-322. [doi: 10.1093/her/cyr005] [Medline: 21321008]

29. Semplonius T, Willoughby T. Long-term links between physical activity and sleep quality. Med Sci Sports Exerc 2018 Dec;50(12):2418-2424. [doi: 10.1249/MSS.0000000000001706] [Medline: 30048409]
30. Borodulin KM, Evenson KR, Wen F, Herring AH, Benson AM. Physical activity patterns during pregnancy. Med Sci Sports Exerc 2008 Nov;40(11):1901-1908 [FREE Full text] [doi: 10.1249/MSS.0b013e31817f1957] [Medline: 18845974]
31. Temple VA, Walkley JW. Physical activity of adults with intellectual disability. J Intellec Dev Disabil 2009 Jul 10;28(4):342-353. [doi: 10.1080/13668250310001616380]
32. Adadi A, Berrada M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access 2018;6:52138-52160. [doi: 10.1109/access.2018.2870052]

Abbreviations

ACM: Association for Computing Machinery
ERIC: Education Resources Information Center
MET: metabolic equivalent
PA: physical activity
PAL: physical activity level
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RQ: research question

Edited by C Lovis; submitted 17.07.22; peer-reviewed by B Hoyt; comments to author 18.09.22; revised version received 25.11.22; accepted 27.11.22; published 06.03.23.

Please cite as:
Diaz C, Caillaud C, Yacef K
Mining Sensor Data to Assess Changes in Physical Activity Behaviors in Health Interventions: Systematic Review
JMIR Med Inform 2023;11:e41153
URL: https://medinform.jmir.org/2023/1/e41153
doi:10.2196/41153
PMID:36877559

Review

Methods Used in the Development of Common Data Models for Health Data: Scoping Review

Najia Ahmadi[1], MSc; Michele Zoch[1], Dipl-Wi-Inf; Patricia Kelbert[2], MSc; Richard Noll[3], MSc; Jannik Schaaf[3], Dr rer med; Markus Wolfien[1,4], Dr-Ing; Martin Sedlmayr[1], Prof Dr

[1]Institute for Medical Informatics and Biometry, Carl Gustav Carus Faculty of Medicine, Technische Universität Dresden, Dresden, Germany

[2]Fraunhofer Institute for Experimental Software Engineering IESE, Kaiserslautern, Germany

[3]Institute of Medical Informatics, Goethe University Frankfurt, University Hospital, Frankfurt, Germany

[4]Center for Scalable Data Analytics and Artificial Intelligence, Dresden/Leipzig, Germany

Corresponding Author:
Najia Ahmadi, MSc
Institute for Medical Informatics and Biometry
Carl Gustav Carus Faculty of Medicine
Technische Universität Dresden
Fetscherstr 74
Dresden, 01307
Germany
Phone: 49 351458 87 7704
Email: najia.ahmadi@tu-dresden.de

Abstract

Background: Common data models (CDMs) are essential tools for data harmonization, which can lead to significant improvements in the health domain. CDMs unite data from disparate sources and ease collaborations across institutions, resulting in the generation of large standardized data repositories across different entities. An overview of existing CDMs and methods used to develop these data sets may assist in the development process of future models for the health domain, such as for decision support systems.

Objective: This scoping review investigates methods used in the development of CDMs for health data. We aim to provide a broad overview of approaches and guidelines that are used in the development of CDMs (ie, common data elements or common data sets) for different health domains on an international level.

Methods: This scoping review followed the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist. We conducted the literature search in prominent databases, namely, PubMed, Web of Science, Science Direct, and Scopus, starting from January 2000 until March 2022. We identified and screened 1309 articles. The included articles were evaluated based on the type of adopted method, which was used in the conception, users' needs collection, implementation, and evaluation phases of CDMs, and whether stakeholders (such as medical experts, patients' representatives, and IT staff) were involved during the process. Moreover, the models were grouped into iterative or linear types based on the imperativeness of the stages during development.

Results: We finally identified 59 articles that fit our eligibility criteria. Of these articles, 45 specifically focused on common medical conditions, 10 focused on rare medical conditions, and the remaining 4 focused on both conditions. The development process usually involved stakeholders but in different ways (eg, working group meetings, Delphi approaches, interviews, and questionnaires). Twenty-two models followed an iterative process.

Conclusions: The included articles showed the diversity of methods used to develop a CDM in different domains of health. We highlight the need for more specialized CDM development methods in the health domain and propose a suggestive development process that might ease the development of CDMs in the health domain in the future.

(JMIR Med Inform 2023;11:e45116) doi:10.2196/45116

KEYWORDS

common data model; common data elements; health data; electronic health record; Observational Medical Outcomes Partnership; stakeholder involvement; Data harmonisation; Interoperability; Standardized Data Repositories; Suggestive Development Process; Healthcare; Medical Informatics;

Introduction

Rationale

Integration of heterogeneous data is a ubiquitous topic in modern medicine. The arising large variety of data has the potential to provide in-depth insights about different aspects of clinical care and can lead to improvements in health care [1,2]. Yet, challenges, such as the identification and access of relevant data, the association between different data sources, and the assurance of data quality given the structural variations among data sources, still pose major barriers [3,4]. Common data models (CDMs) provide the possibility of harmonizing data from disparate sources, storing information in a standard structure by defining the syntax and semantics of data, and enabling operations on data using standard analysis methods [5]. In particular, a CDM contains a unified set of metadata, allowing data and its information content to be shared across applications and institutional borders, and thus enabling harmonized data integration and analysis on an international scale [6].

In the health domain, there are different types of CDMs (eg, CDMs for harmonization and storage of electronic health record–based patient data). An example is the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) developed by the Observational Health Data Science and Informatics (OHDSI) community, which ensures homogeneous storage of observational health care data across different databases with similar formats and terminologies [7]. There are also further CDMs for clinical data, like Sentinel CDM, Clinical Data Interchange Standards Consortium (CDISC) Study Data Tabulation Model (SDTM), and National Patient-Centered Clinical Research Network (PCORnet) [8], and data warehouse models, like Informatics for Integrating Biology and the Bedside (i2b2) [9]. Moreover, some CDMs define the data from patient cohorts and describe a medical specialty or a group of diseases. For example, there are specific CDMs for the domain of rare diseases [10,11] or radiology [12]. Overall, there is a large variety of CDMs in the literature for common, rare, and context-specific medical examinations, and each of them follows a more self-defined development process.

As described by Melles et al [13], a practical design meets the users' needs. While designing a CDM in the health domain, in addition to the developers (ie, IT staff and computer scientists), the primary stakeholders (ie, patients and clinicians) are particularly interested in the outcome. It is therefore recommended to include them in the design process as early as possible [13,14]. In addition to the stakeholders, the medical context is also quite complex and requires extensive medical and technical expertise to ensure the usefulness of the model after its development. This is why the development process of a CDM is critical and a comprehensive development method or guideline is necessary.

Studies, such as those by Gericke and Blessing [15] and Bobbe et al [16], have already tried to determine the commonalities and differences in development processes across disciplines. Bobbe et al [16] performed a comparison of design models from academic theory and professional practice, and discussed 8 types of design processes. In particular, the basic design cycle, V design process, human-centered design, hypercyclic design, Munich procedural model, double diamond model, frog model, and IDEO model were presented. Additionally, Melles et al [13] introduced categories for models, namely, whether a model is activity-based or stage-based, solution-oriented or problem-oriented, and design-focused or project-focused.

However, given the complexity of the health domain and the importance of many stakeholders taking part in the process, it might be difficult to transfer models from other disciplines. This is why we aim to derive such a process and review the available CDM instances in the domain. Exemplarily, the results of this scoping review will be integrated into the design and development of a CDM for the SATURN ("Smartes Arztportal für Betroffene mit unklarer Erkrankung" ["Smart physicians' platform for patients with unclear diseases"]) Project in the future [17]. This project aims to develop an artificial intelligence–based diagnosis support tool for primary care physicians. With the help of user-centered design, the requirements of a decision support tool, especially for noncharacteristic symptoms, will be studied. The medical focus is on the diagnosis of unclear and rare medical conditions. This is why, in this review, we focus on the similarities between the CDM development methods in rare medical conditions and common medical conditions in order to determine whether the methods for common medical conditions can be adopted for rare medical conditions as well. On a technical level, rule-based systems, machine learning, and case-based reasoning will be implemented. As part of this project, CDMs for 3 groups of rare diseases, namely, endocrinology, gastroenterology, and pneumology, will be developed.

Our review contributes to the analysis of CDM development methods in the health domain on an international scale and aims to explore the actual involvement of stakeholders, especially medical experts, in the development process. To the best of our knowledge, this is the first scoping review focusing on CDM development methods in the health domain.

Objectives and Research Questions

This scoping review has been conducted to provide an overview of the methods used for the initial and further development of CDMs in the health domain. We divided the overall development process into conception, users' needs collection (eg, collection of evidence, review of the literature, and guidelines), and implementation, as well as individual evaluations within the phases. We consider the conception phase as an initial step, where the CDM is theoretically designed along with stakeholders. Subsequently, the essential elements previously identified are gathered in the "users' needs collection" phase. The finalized process, in which the conceptualized model is implemented and ready-to-use, is termed the implementation phase.

According to the rationale and objective explained above, this scoping review examines the following questions:

1. How are CDMs methodically developed in the health domain? What requirement analysis methods, design processes, and validation methods were used?

2. How or when do stakeholders, especially medical experts, get involved in the development process?
3. How can the CDM development methods be classified based on their requirement analysis methods, design processes, validation methods, and model type?

Methods

Protocol and Registration

To ensure methodological quality, this scoping review has followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) checklist [18]. According to this checklist, we published and registered the review protocol [19]. Out of the 22 items of the PRISMA checklist, 20 have been considered in this review (Multimedia Appendix 1).

Search Strategy

To achieve a comprehensive query, an initial search was performed in PubMed with the term "common data model." Six randomly chosen articles matching the topic were analyzed [10-12,20-22]. The keywords associated with the articles listed in Table 1 were considered and subsequently tested in the query. The combination of terms that delivered the highest number of matching articles was included in our final search string.

Some studies used the term *data set* [11], and others defined alternative data elements that can be part of a data set or data model [10]; thus, to avoid the exclusion of certain studies, we jointly used the following terms in our search string: *common data model*, *common data element*, and *common data sets*. We also added the short forms of these terms in our search string and analyzed the relevance of the results by simply looking into the resulting literature. Additionally, we added the following terms in our search string to ensure that the included CDMs were developed within the health domain: medical, medicine, health, healthcare, health care, electronic health, clinical, and disease. The search string used in PubMed is presented in Table 2. It was developed as a combination of the mentioned terms, their possible variations, and where applicable, Medical Subject Headings (MeSH) [23]. The search strings used in the other 3 databases have been provided in Multimedia Appendix 2.

The query was designed and tested by the author NA and was approved by all coauthors. The resulting articles were added to Rayyan (Rayyan Systems Inc) [24] for further screening and annotation.

Table 1. Six randomly chosen articles for the construction of the search string and their keywords.

Article title	Keywords
The EPIRARE proposal of a set of indicators and common data elements for the European platform for rare disease registration [10]	Registries, common data elements, European platform, rare diseases, patient registration, and EPIRARE
A methodology for a minimum data set for rare diseases to support national centers of excellence for healthcare and research [11]	Common data elements, interoperability, metadata, minimum data set, national health program, and rare diseases
Development and validation of the Radiology Common Data Model (R-CDM) for the international standardization of medical imaging data [12]	*Metadata*, standardization, and radiology information system
Common data model for natural language processing based on two existing standard information models: CDA+GrAF [20]	Natural language processing, medical informatics, data model, information model, HL7 clinical document architecture, and ISO graph annotation format
Genomic common data model for biomedical data in clinical practice [21]	High-throughput nucleotide sequencing, data analysis, and observational study
Towards a newborn screening common data model: The Utah Newborn Screening Data Model [22]	Newborn screening, newborn screening laboratory information management system, common data model, interoperability, electronic data exchange, NBS, LIMS, and standards

Table 2. Search strings used to identify articles from PubMed.

Search aspects	Variations	Search string[a]
Common data model	Common data model (CDM), common data element (CDE), and common data sets (CDS)	("common data model" AND CDM) OR ("common data element*" AND CDE) OR "Common Data Elements"[Mesh] OR "common dataset*" OR "common data set*"
Health care	Medical, medicine, health, healthcare, health care, electronic health, and disease	medical OR medicine OR "Medicine"[Mesh] OR health OR "Health"[Mesh] OR healthcare OR "health care" OR "electronic health" OR clinical OR disease OR "Disease"[Mesh]

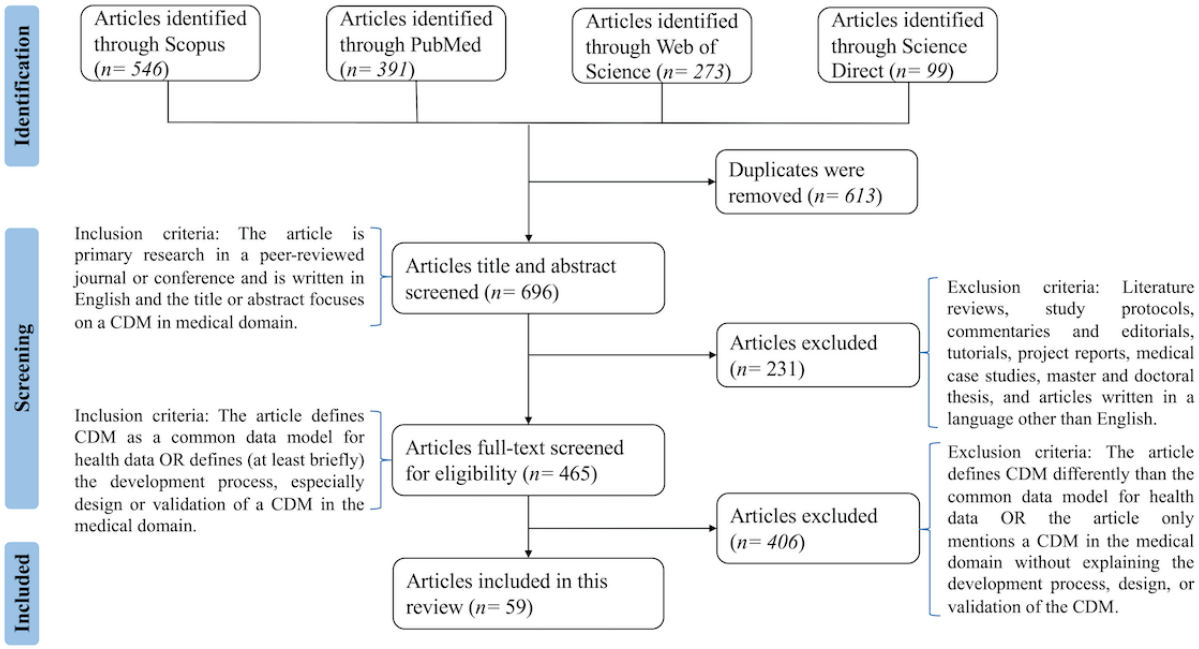
[a]The common data model and health care search terms were combined with "AND."

In particular, literature from 2000 to 2022 was considered, which is an extension of the previously published study protocol [19]. It is also noteworthy that the MeSH terms were only available in PubMed. The language of the articles was limited to English. Using the Boolean operators "AND" and "OR," the systematic search was carried out in the following electronic databases: PubMed, Web of Science, Science Direct, and Scopus. The search was performed in March 2022. The publication date tag in PubMed and Web of Science was set to January 1, 2000, to March 15, 2022, and that in Science Direct and Scopus was set to 2000 to 2022 (it is not possible to specify the month and day in Science Direct and Scopus).

Inclusion and Exclusion Criteria

The inclusion and exclusion criteria are summarized in Multimedia Appendix 3 and are visualized along with the number of outcome articles in Figure 1.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart showing the paper selection process and the inclusion and exclusion criteria. CDM: common data model.



Selection and Review of Articles

Duplicates were removed using the built-in function in Rayyan [24]. The process of deletion was monitored by the author NA. After eliminating duplicates, the selection of studies was performed in 2 steps. The title and abstract screening steps were performed by the authors in groups of two. The articles were tagged as “include,” “exclude,” or “maybe.” Tagged articles were decided upon based on the tags described in Table 3. Disagreements were resolved by a third author. This process was initially carried out on 10% of the articles to confirm the accuracy of our inclusion and exclusion criteria, and clarify ambiguities. After the title and abstract screening, the full text of the included articles was screened by the authors, again in groups of two. The selected articles were included in the data extraction step.

Table 3. Description of tags used by the authors in the article screening process.

Author 1	Author 2	Decision
Include	Include	Included
Include	Exclude	Discuss and decide together
Include	Maybe	Include
Exclude	Maybe	Exclude
Maybe	Maybe	Discuss and decide together

Data Charting and Extraction Process

A data charting table was developed and refined throughout the study, with several iterations. This table contained a list of items that were extracted from all included publications. All authors examined 10% of the articles for the defined data items and refined the data charting table, if necessary. The data charting table, including the extracted information from articles, is included in Multimedia Appendix 4.

For each article, we focused on 4 major aspects: (1) the meta information, such as DOI, authors, year, country, and project name, if applicable; (2) the medical condition for which the CDM was built, whether the condition is rare or common, the organ affected by the condition, and whether the condition is long term (longer than a year) or short term; (3) methodological information, such as requirement analysis, design, and validation process; whether the design process was linear or iterative; and advantages and disadvantages of the method, as stated in the respective article; and (4) information about stakeholder involvement. The extracted data elements, their categories, and their definitions are shown in Table 4.

Table 4. Data extraction sheet with specified elements, categories, and subcategories, including their definitions.

Category and subcategory	Definition
Meta information	
DOI	A link to the article
Author	First author's name
Publication year	Year of the publication date of the article
Country of study	Country of the leading author's affiliation
Project name	If applicable; when the CDM[a] study was part of a project/consortium
Medical background	
Medical condition	Name of the medical condition for which the CDM was built
Organ function	Organ affected by the medical condition
Short-term/long-term condition	Short term: less than a year; long term: longer than a year
Is the condition rare or common?	Is the medical condition considered rare or common based on its occurrence? Available answers: common medical condition, rare medical condition, and conditions that can be rare and common.
Requirement analysis method	
Literature analysis	It includes searching in a variety of literature, such as extraction of frequent CDEs[b] from real-world data, data harmonization across studies, multicenter longitudinal and observational studies, consensus documents and guidelines, primary outcome data of trials, review of instruments, and forms like report forms, users' needs collection forms, etc.
Interview/questionnaire	It includes expert interviews, focus group meetings, working group meetings, consensus meetings, workshops and discussions, and online surveys.
Delphi	Delphi or modified Delphi was used. Delphi techniques involve experts evaluating complex issues iteratively, where knowledge is incomplete or uncertain. Typically, the response from the previous questionnaire is appended to the next questionnaire [25].
Review of existing CDEs	When an existing CDE was validated/reviewed.
Design	
Creation of new CDEs	If there were no CDEs in the domain and the experts tried to come up with some CDEs using literature in the field.
Modification of existing CDEs	If existing CDEs in a disease domain were modified.
Reuse of existing CDEs (without modification)	If existing CDEs in the domain were used without any modification.
Validation	
External experts	It includes only external validation of any sort, such as public reviews on a website from experts or nonexperts in the field. Excluded are experts that were part of the conception process of the model.
Others	Any other type of validation, such as internal reviews, working group consensus, etc.
Model type	
Iterative	When at least one iterative process was performed during development of the CDM.
Linear	When there was no iteration in the process.
Stakeholder information	
Were stakeholders involved in the design process?	Yes/no
Which stakeholders were involved?	Patients' representatives, clinicians, domain experts, computer scientists, IT personnel, and registry staff
When did they get involved in the process?	In users' needs collection (when experts were involved in the preanalysis step, eg, collection of evidence, review of literature, guidelines, etc), in conception (when experts were involved in conception of the CDEs), in evaluation (when the model was evaluated via experts), and in implementation (when experts were involved in the implementation of the model).
What was the nature of stakeholder involvement?	Through expert workshops, semistructured interviews, questionnaires, etc
Pros and cons of methods as mentioned in the article	

Category and subcategory	Definition
Pros	Advantages of the method as stated in the article
Cons	Disadvantages of the method as stated in the article

[a]CDM: common data model.

[b]CDE: common data element.

Visualization and Summarization of Results

At the end of the data extraction, the data items collected in Table 4 were summarized and visualized. A flowchart according to the PRISMA-ScR guidelines was designed to show the article processing approach (Figure 1). Tables, timeline plots, histogram charts, pie plots, and scatter histograms were used to display the extracted data items. The graphics and the required analysis were performed using Python version 3.9.12 (Python Software Foundation), with matplotlib, pandas, and NumPy packages. The script used for the plots is publicly available [19,26].

First, we aimed for a broad overview of available CDMs and whether original CDMs were developed or existing CDMs were modified, as well as whether they focused on common or rare diseases and addressed a specific organ function. Second, to answer our first research question, we documented the medical domain of each article, whether the medical condition was considered as long term (more than a year) or short term (less than a year), and the affected organ as stated in the respective original article. To classify the development process of the CDMs, we documented 4 categories of data information for each article: requirement analysis, design, validation, and model type (Table 4). We categorized the methodology that was used for the requirement analysis (ie, why a CDM was needed), as well as the context to design a set of common data elements (CDEs). For validation, we distinguished between external evaluation and any other type of evaluation. The "other" category included the evaluations performed by the same clinical experts who were involved in the conception process, such as working group consensus, user evaluations, reviews performed via the members of the project, statistical tests, and pilot tests conducted within the project. Additionally, we investigated stakeholder involvement in the development stages in those studies and whether the studies followed an iterative or linear method of development. We used the advantages and disadvantages of the methods as stated in the articles (Table 4) and formulated them into a list of constraints in the area of CDM development to further highlight the need for streamlined methods. Finally, after analyzing the included CDMs, we summarized the most frequent methods used in the included literature in a suggestive development process that could be a reasonable basis to start with when developing a novel model.

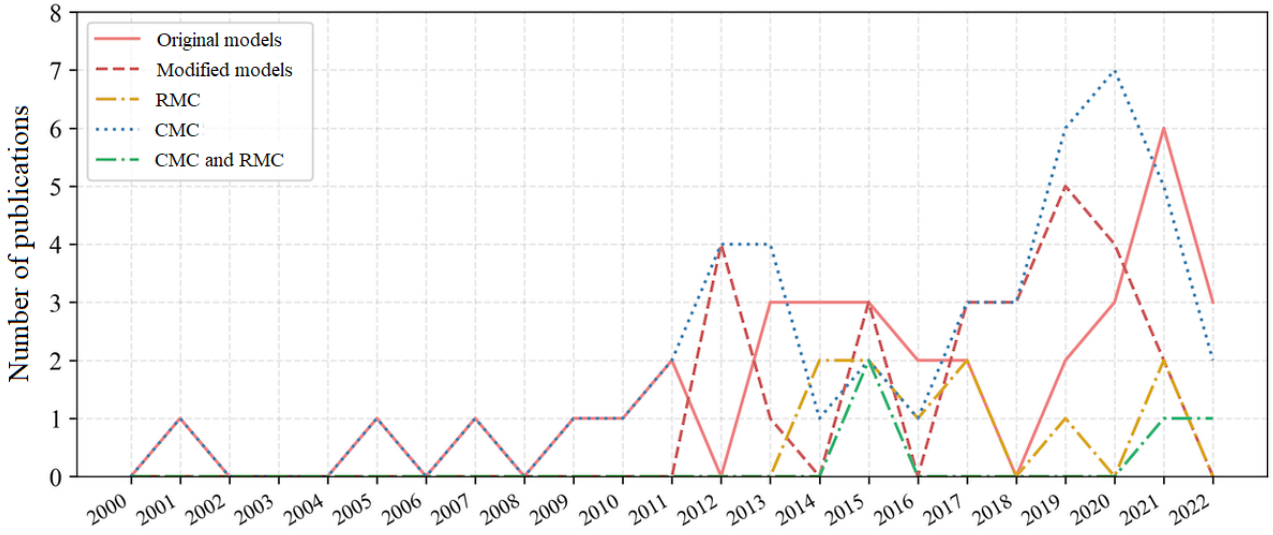
Results

Selection of Articles

In total, we identified 1309 articles from PubMed, Web of Science, Science Direct, and Scopus search engines. From the identified articles, after duplicate removal, 695 articles were included in the title and abstract screening. Finally, 465 articles underwent full-text screening, and of these, 59 matched the full-text screening criteria of this review and were finally included. We excluded articles that did not describe the development or evaluation of a CDM in the health domain. Additionally, articles that were not publicly available and those in a language other than English were excluded. The article identification process along with the inclusion and exclusion criteria are shown in Figure 1.

The selected articles defined CDMs, common data sets, or CDEs for common or rare medical conditions. All included articles were published between 2000 and 2022. As shown in Figure 2, the number of articles that focused on CDM development increased after 2011 and continued to increase in the last years.

Figure 2. The number of publications focusing on common data model (CDM) development per year from 2000 to 2022. The line chart compares the number of articles developing original CDMs (original models) with the number of articles developing CDMs via modification of existing models (modified models), and compares the number of articles developing CDMs for rare medical conditions (RMCs), the number of articles developing CDMs for common medical condition (CMCs), and the number of articles developing CDMs for both kinds of conditions (CMCs and RMCs). In addition to the increase in the number of articles from 2011 in general, we can see that CDMs for rare diseases were only developed starting from 2014.



Country of Publication

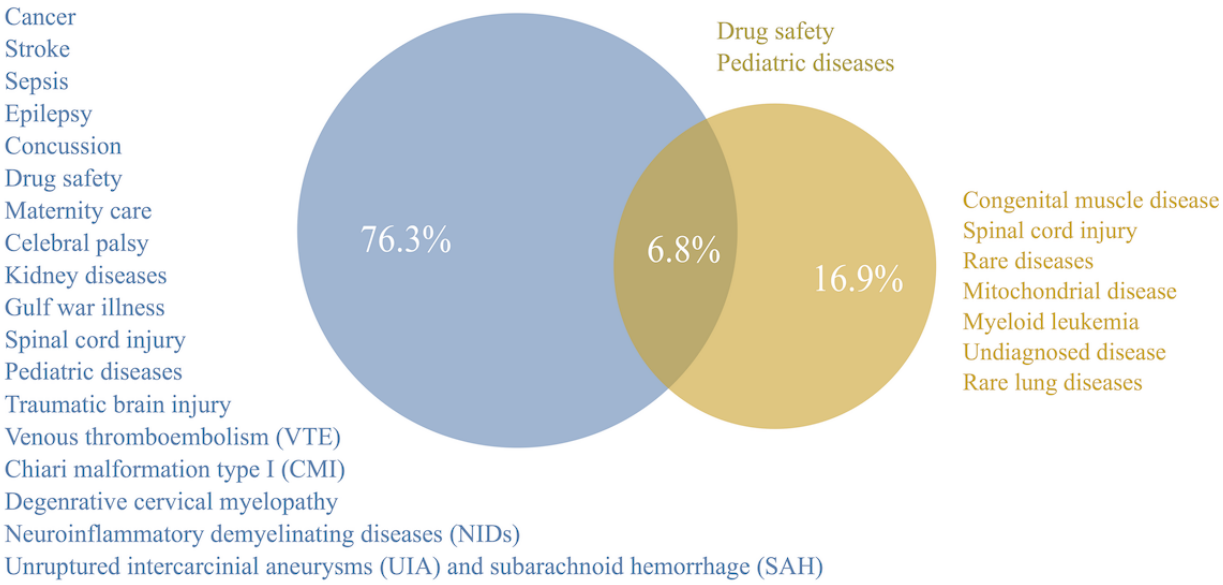
We categorized the articles into countries based on the affiliation of the first author. Among the 59 articles, 26 (44%) were published in the United States, 8 (14%) were published in Canada, and 6 (10%) were published in Germany. The number of articles according to country is as follows: Belgium, 2 [27,28]; Canada, 8 [29-36]; China, 1 [37]; Denmark, 2 [38,39]; France, 2 [11,40]; Germany, 6 [41-46]; Italy, 1 [10]; Spain, 1 [47]; Republic of Korea, 1 [48]; Norway, 3 [49-51]; Switzerland, 1 [52]; Taiwan, 1 [53]; the Netherlands, 1 [54]; United Kingdom, 3 [55-57]; and United States, 26 [58-83].

Medical Conditions and Their Domains

According to our research, CDMs were developed for a variety of medical domains in the past 22 years; however, we divided them into 3 categories, namely, rare, common, and rare and common (both). An aggregated list of the medical conditions and their domains is shown in Figure 3. A full list of the medical conditions extracted during this scoping review is shown in Multimedia Appendix 4. An organ function overview and the long- and short-term conditions are shown in Multimedia Appendix 5. Among these, 10 (17%) CDMs were designed for rare medical conditions, such as myeloid leukemia and rare lung diseases, and mitochondrial diseases [41,44-46,59]. Moreover, 1 CDM, namely, the CDM in the study by Berger et al [44], was designed for undiagnosed diseases in general.

Among the 59 articles, 45 involved the development of a CDM for common medical conditions. These included traumatic brain injury [27,28,30], spinal cord injury in children and youth [67], dental caries [68], sport-related concussion [65], cerebral palsy [29], degenerative cervical myelopathy [55], unruptured intracranial aneurysms and subarachnoid hemorrhage [32,42,55,60], Chiari malformation type I [63], breast implant [43], stroke [37], venous thromboembolism [33], pediatric epilepsy [61], pediatric critical illness [62], pregnancy drugs and treatments [49], sepsis [31], medication use in pregnancy and breastfeeding [40], degenerative cervical myelopathy [55], Gulf War illness [58], neuroinflammatory demyelinating disease [43], traumatic brain injury [27], and neurologic disorder and stroke [69]. Wandner et al [66] focused on clinical pain management, and Jaboyedoff et al [52] focused on pediatric diseases in general.

Figure 3. Characteristics of the included studies. A Venn diagram showing the proportions of identified common data models (CDMs) for common medical conditions (76.3%; blue), rare medical conditions (16.9%; golden yellow), and medical conditions that could fit into both categories (6.8%). Additionally, an aggregated list of medical conditions that CDMs were developed for in the studies is shown in 3 different colors according to their categories.



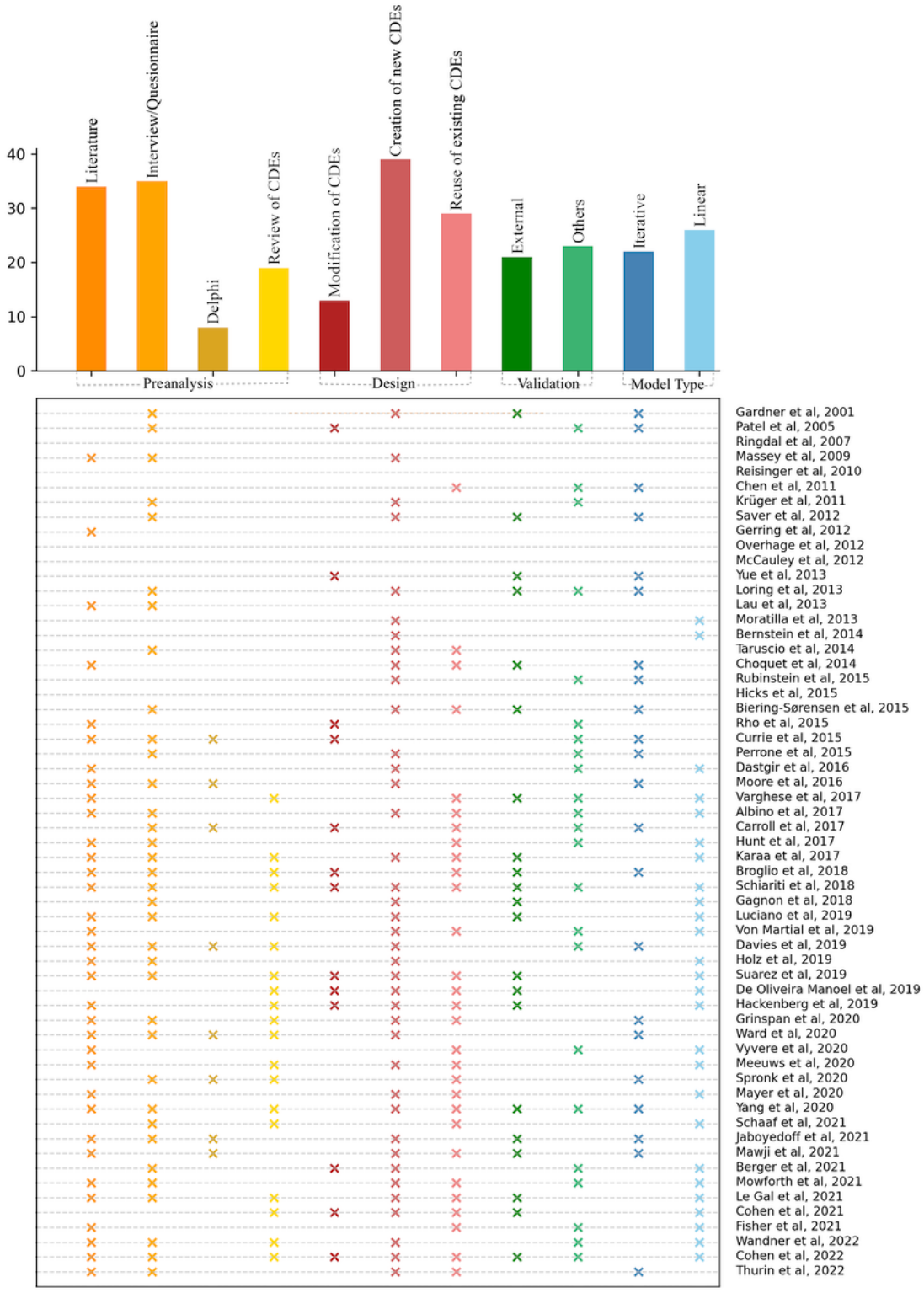
Stakeholder Involvement

To investigate the involvement of stakeholders, we summarized at which particular stage they were involved in the CDM development process. Out of the 59 included articles, 54 (92%) mentioned at least one stakeholder in the design process. Additionally, we were interested in the different types of stakeholders that were involved, how they were involved, and at what stage of the process they typically got involved. As shown in Figure S1 in Multimedia Appendix 6, stakeholders were mostly involved in the initial stage, namely, the conception phase. Domain experts and clinicians were the most common stakeholders involved in the studies (Figure S2 in Multimedia Appendix 6). Additionally, while many different methods were used to involve the stakeholders, such as expert groups, surveys, consensus meetings, interviews, teleconferences, questionnaires, and workshops, "working group" was the most frequent method used (Figure S3 in Multimedia Appendix 6).

Design Process

The methods used in the articles for designing a CDM were literature analysis, interview, Delphi, and review of existing CDEs. From our extraction table (Multimedia Appendix 4), we noted that 39 articles involved the definition of an original model/set of CDEs, 13 involved the modification of an existing set of CDEs, and 29 involved the use of an existing set of CDEs without any modifications. The external evaluation included web-based feedback, public review and comments, and feedback in a conference, among others. Finally, we found that 26 articles involved a rather linear design method and 22 others involved an iterative process. The list of articles that involved the use of each of these categories is shown in Figure 4. Detailed information is presented in Multimedia Appendix 4.

Figure 4. Methodological information on the articles [10,11,27-83]. The y-axis shows the list of articles by publication year. The x-axis shows the methodological categories. The scatter plot includes a cross mark when the Boolean is true for a specific article, for example, if the authors have used literature analysis as a preanalysis method, a cross (x) is added. The sum of cross marks in each column contributes to the bar size of the bar plot positioned on the x-axis. To improve visibility, each subcategory is shown with a different color. The subcategories of the same category are grouped via the same family of colors. CDE: common data element.

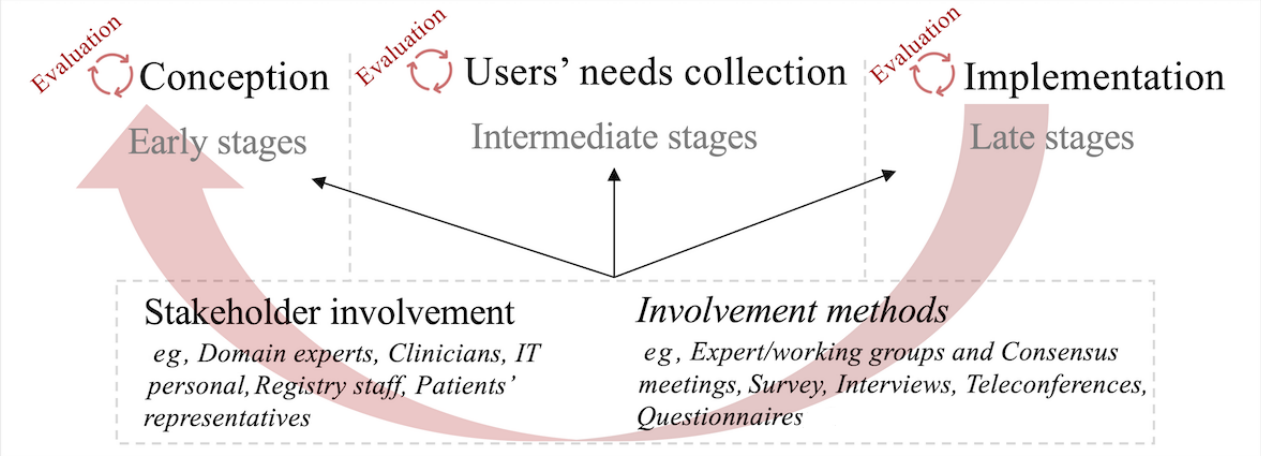


Methodological Constraints Highlighted in Previous Studies

The included articles presented a range of constraints in the development process from the methods used in the different stages of the process to the applicability of the outcome elements. For example, Thurin et al [40] performed interviews with a single data access provider per data source and mentioned that other data access providers might conceptualize the data source differently. Additionally, they tested the applicability of the developed model only on the included data sources in the project. The model might require modification to use it with

other data sources. The limited sample size used to test the developed model is a common problem in rare conditions [44] given the rarity of the disease. One of the limitations mentioned by Broglio et al [65] is that some of their developed CDEs require special expertise that might not be implementable in certain settings. Grinspan et al [61] mentioned that some subcategories of epilepsy syndrome were merged at a level higher into a single category, which might have led to reduced data resolution, although uphill mapping is often used, especially in the OMOP context [5]. Additionally, the elements considered do not cover every possible influencing element, and the source was limited to only US-based patients, which means the elements can differ once an international data level is considered. They also included CDEs that were documented as free text, and processing of such elements might require natural language processing applications. The authors also highlighted the possible bias caused by the methodology used for consensus and discussion, and the Delphi approach, focus groups, and interviews might have also influenced the outcome of the study.

Figure 5. Summary of a basic common data model development process.



Essence of the CDM Development Process

Our outcomes showed that a heterogeneous variety of methods or processes were used in CDM development in the included articles, which highlights the need for a more streamlined field-specific development method. Therefore, we summarized our analysis outcomes into a suggestive development process (Figure 5), considering the 3 development steps that have been identified from the included models in this study, namely, conception, users' needs collection, and implementation. We suggest that evaluation and validation should be integrated into every stage of development, which gives the stages an iterative nature, and feedback should be integrated into the process as much as possible. We also emphasize the involvement of stakeholders in the process as early as possible and propose continuous involvement until the end of the development process because in every phase, questions might arise that need to be answered from different perspectives.

Discussion

Overview

One of the major challenges faced by CDM developers in the health domain is the lack of a comprehensive methodology or workflow to follow, which is also reflected in this review. The general models from industrial design and even academia (eg, the model introduced by Bobbe et al [16]) do not generally translate one-to-one to the health domain. The medical context is usually complex, and the involvement of stakeholders, such as clinicians, patients' representatives, and IT staff, is of utmost importance to ensure the applicability of a to-be-developed CDM. In addition, user-friendly, adaptable, and straightforward models are preferred in health care as one can start working with them without requiring a substantial amount of time [84].

This scoping review provides a summary of the development methods for CDMs and categorizes them based on the requirement analysis method, design process, validation approach, and model type. A variety of methods were used in the requirement analysis step in the articles, starting from searching in different types of literature and medical guidelines [43,44] to interviews [29], the Delphi approach [31], and a review of existing CDEs. A full list of these articles is shown in Figure 4 and Multimedia Appendix 4.

The majority of the developed CDMs have been designed for common medical conditions, and only 10 articles involved the design of a particular CDM for rare diseases. However, we did not find a significant difference in the development process of a CDM for rare and common conditions. Interestingly, based on our analysis, we can conclude that common medical conditions were the focus of CDM studies from early 2000, whereas the first CDM for rare conditions was developed in 2014. Despite methodological similarities, every article usually mentioned following a more individualistic method of development. This may arise because rare conditions occur rarely and the number of patients included in studies is limited [44]. Moreover, finding an expert for each rare or unclear disease is a challenging task. Additionally, most of the information crucial in the diagnosis of such diseases (like symptoms or phenotypes and genotypes) is currently stored in unstructured forms (eg, clinical notes). Extraction of such information requires a lot of time and effort from technical and clinical stakeholders [41].

Thus, given the variety of studies, the methods used for common conditions might be adaptable for rare conditions. Considering that a CDM is an essential part of data harmonization (a necessity in the health domain), we see highly emphasized development models as essential. Therefore, after analyzing the included CDMs, we summarized a suggestive development process that is shown in Figure 5, which could be the starting point for conceptualizing and implementing novel CDMs.

Limitations

The findings of our study are subject to certain limitations. First, our analysis is restricted to the selected databases, namely, PubMed, Web of Science, Science Direct, and Scopus. Additionally, the scope of our investigation is confined to articles published within a specific time frame and written in English. Moreover, we did not conduct any assessment of the quality of the included articles. In addition, it may also be worth noting that the authors of this review have varying interdisciplinary backgrounds, expertise levels, and experiences in the CDM field. However, to optimize the screening and analyzing processes, we performed them in pairs and first tested the method on a subset of 10% of the articles, resulting in a minimal number of conflicts.

Conclusion

We considered 4 steps in the development of a CDM: conception, users' needs collection, implementation, and evaluation. We could identify 4 groups of methods that were most often used in the articles as part of the requirement analysis of the CDM development process. These were literature analysis, interviews, Delphi approaches, and review of existing CDEs. The articles considered in this review either developed a new CDE or made use of an existing set of CDEs with or without modification.

Most of the articles involved at least one stakeholder from among domain experts, clinicians, IT staff, registry staff, and patients' representatives, and mostly from the initial step, which was conception. The methods used to involve the stakeholders were expert groups, surveys, consensus meetings, interviews, working groups, teleconferences, questionnaires, and workshops, and among these, working groups were most often used.

We conclude that the methods used in the development of CDMs in the health domain are heterogeneous and this field is lacking solid guidelines that may ease up this process, especially in terms of the reusability and adaptability of a CDM. This is why the proposed outline (Figure 5) could be a reasonable basis to start with. In our future work, we plan to test and improve the proposed outline for developing a CDM.

Acknowledgments

This work was accomplished as part of the SATURN (Smartes Arztportal für Betroffene mit unklarer Erkrankung; Smart physicians' platform for patients with unclear diseases) Project funded by the German Federal Ministry of Health as part of the research focus "Digital Innovation," Module 3: "Smart Algorithms and Expert Systems" (funding codes: 2520DAT02C, 2520DAT02B, and 2520DAT02D).

Data Availability

The script used for analysis and visualization in this review is available at GitHub [26], and the study protocol can be accessed on the Open Science Framework (OSF) [19].

Authors' Contributions

NA, MZ, PK, RN, MW, JS, and MS contributed to conceptualization and methodology. NA contributed to data acquisition. NA, MZ, PK, RN, and MW contributed to the literature screening. NA contributed to data analysis and interpretation. NA contributed to writing and preparing the original draft. NA, MZ, PK, RN, MW, and JS contributed to reviewing and editing the manuscript. NA and MW contributed to the visualization. MS contributed to resources. All authors take responsibility for the scientific integrity of the work. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist.
[PDF File (Adobe PDF File), 940 KB - medinform_v11i1e45116_app1.pdf]

Multimedia Appendix 2
Search strings used in the PubMed, Web of Science, Science Direct, and Scopus databases to search for articles.
[DOCX File , 17 KB - medinform_v11i1e45116_app2.docx]

Multimedia Appendix 3

Inclusion and exclusion criteria for the title and abstract screening, and the full-text screening.
[DOCX File , 14 KB - medinform_v11i1e45116_app3.docx]

Multimedia Appendix 4
Additional study information.
[XLSX File (Microsoft Excel File), 64 KB - medinform_v11i1e45116_app4.xlsx]

Multimedia Appendix 5
Characteristics of the included studies.
[PDF File (Adobe PDF File), 99 KB - medinform_v11i1e45116_app5.pdf]

Multimedia Appendix 6
Summary of derived stakeholder information.
[PDF File (Adobe PDF File), 198 KB - medinform_v11i1e45116_app6.pdf]

References

1. Dash S, Shakyawar SK, Sharma M, Kaushik S. Big data in healthcare: management, analysis and future prospects. J Big Data 2019 Jun 19;6(1):54. [doi: 10.1186/s40537-019-0217-0]
2. Razzak MI, Imran M, Xu G. Big data analytics for preventive medicine. Neural Comput Appl 2020 Mar 16;32(9):4417-4451 [FREE Full text] [doi: 10.1007/s00521-019-04095-y] [Medline: 32205918]
3. Asche CV, Seal B, Kahler KH, Oehrlein EM, Baumgartner MG. Evaluation of healthcare interventions and big data: review of associated data issues. Pharmacoeconomics 2017 Aug;35(8):759-765. [doi: 10.1007/s40273-017-0513-5] [Medline: 28474299]
4. Kent S, Burn E, Dawoud D, Jonsson P, Østby JT, Hughes N, et al. Common problems, common data model solutions: evidence generation for health technology assessment. Pharmacoeconomics 2021 Mar;39(3):275-285 [FREE Full text] [doi: 10.1007/s40273-020-00981-9] [Medline: 33336320]
5. Maier C, Lang L, Storf H, Vormstein P, Bieber R, Bernarding J, et al. Towards implementation of OMOP in a German university hospital consortium. Appl Clin Inform 2018 Jan;9(1):54-61 [FREE Full text] [doi: 10.1055/s-0037-1617452] [Medline: 29365340]
6. Simko LC, Chen L, Amtmann D, Gibran N, Herndon D, Kowalske K, et al. Challenges to the standardization of trauma data collection in burn, traumatic brain injury, spinal cord injury, and other trauma populations: a call for common data elements for acute and longitudinal trauma databases. Arch Phys Med Rehabil 2019 May;100(5):891-898 [FREE Full text] [doi: 10.1016/j.apmr.2018.10.004] [Medline: 31030731]
7. Hripcsak G, Duke J, Shah N, Reich C, Huser V, Schuemie M, et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. Stud Health Technol Inform 2015;216:574-578 [FREE Full text] [Medline: 26262116]
8. Garza M, Del Fiol G, Tenenbaum J, Walden A, Zozus MN. Evaluating common data models for use with a longitudinal community registry. J Biomed Inform 2016 Dec;64:333-341 [FREE Full text] [doi: 10.1016/j.jbi.2016.10.016] [Medline: 27989817]
9. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc 2010 Feb 26;17(2):124-130 [FREE Full text] [doi: 10.1136/jamia.2009.000893] [Medline: 20190053]
10. Taruscio D, Mollo E, Gainotti S, Posada de la Paz M, Bianchi F, Vittozzi L. The EPIRARE proposal of a set of indicators and common data elements for the European platform for rare disease registration. Arch Public Health 2014 Oct 13;72(1):35 [FREE Full text] [doi: 10.1186/2049-3258-72-35] [Medline: 25352985]
11. Choquet R, Maaroufi M, de Carrara A, Messiaen C, Luigi E, Landais P. A methodology for a minimum data set for rare diseases to support national centers of excellence for healthcare and research. J Am Med Inform Assoc 2015 Jan;22(1):76-85 [FREE Full text] [doi: 10.1136/amiajnl-2014-002794] [Medline: 25038198]
12. Park C, You SC, Jeon H, Jeong CW, Choi JW, Park RW. Development and Validation of the Radiology Common Data Model (R-CDM) for the International Standardization of Medical Imaging Data. Yonsei Med J 2022;63(Suppl):S74. [doi: 10.3349/ymj.2022.63.s74]
13. Melles M, Albayrak A, Goossens R. Innovating health care: key characteristics of human-centered design. Int J Qual Health Care 2021 Jan 12;33(Supplement_1):37-44 [FREE Full text] [doi: 10.1093/intqhc/mzaa127] [Medline: 33068104]
14. Sacristan JA, Aguaron A, Avendaño C, Garrido P, Carrion J, Gutierrez A, et al. Patient involvement in clinical research: why, when, and how. Patient Preference and Adherence 2016 Apr:631-640. [doi: 10.2147/ppa.s104259]
15. Gericke K, Blessing L. An analysis of design process models across disciplines. In: DS 70: Proceedings of DESIGN 2012, the 12th International Design Conference. 2012 Presented at: 12th International Design Conference; May 21-24, 2012; Dubrovnik, Croatia.

16. Bobbe T, Krzywinski J, Woelfel C. A comparison of design process models from academic theory and professional practice. In: DS 84: Proceedings of the DESIGN 2016 14th International Design Conference. 2016 Presented at: 14th International Design Conference; May 16-19, 2016; Dubrovnik, Croatia.
17. SATURN Projekt. URL: https://www.saturn-projekt.de/ [accessed 2023-07-06]
18. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. Ann Intern Med 2018 Oct 02;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
19. Ahmadi N, Zoch M, Kelbert P, Noll R, Schaaf J, Wolfien M, et al. Methods used in the development of Common Data Models for health data – A Scoping Review Protocol. OSF. URL: https://osf.io/rza84/ [accessed 2023-07-06]
20. Meystre SM, Lee S, Jung CY, Chevrier RD. Common data model for natural language processing based on two existing standard information models: CDA+GrAF. J Biomed Inform 2012 Aug;45(4):703-710 [FREE Full text] [doi: 10.1016/j.jbi.2011.11.018] [Medline: 22197801]
21. Shin S, You S, Roh J, Park Y, Park R. Genomic Common Data Model for Biomedical Data in Clinical Practice. Stud Health Technol Inform 2019 Aug 21;264:1843-1844. [doi: 10.3233/SHTI190676] [Medline: 31438371]
22. Jones D, Shao J, Wallis H, Johansen C, Hart K, Pasquali M, et al. Towards a newborn screening common data model: the Utah Newborn Screening Data Model. Int J Neonatal Screen 2021 Oct 27;7(4):70 [FREE Full text] [doi: 10.3390/ijns7040070] [Medline: 34842615]
23. Chapman D. Advanced search features of PubMed. J Can Acad Child Adolesc Psychiatry 2009 Feb;18(1):58-59 [FREE Full text] [Medline: 19270851]
24. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev 2016 Dec 05;5(1):210 [FREE Full text] [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
25. Niederberger M, Spranger J. Delphi technique in health sciences: a map. Front Public Health 2020;8:457 [FREE Full text] [doi: 10.3389/fpubh.2020.00457] [Medline: 33072683]
26. Ahmadi N. VisualisationWithPython. GitHub. URL: https://github.com/NajiaAhmadi/VisualisationWithPython [accessed 2023-07-06]
27. Meeuws S, Yue JK, Huijben JA, Nair N, Lingsma HF, Bell MJ, et al. Common Data Elements: Critical Assessment of Harmonization between Current Multi-Center Traumatic Brain Injury Studies. J Neurotrauma 2020 Jun 01;37(11):1283-1290 [FREE Full text] [doi: 10.1089/neu.2019.6867] [Medline: 32000562]
28. Vande Vyvere T, De La Rosa E, Wilms G, Nieboer D, Steyerberg E, Maas A, CENTER-TBI ParticipantsInvestigators. Prognostic validation of the NINDS common data elements for the radiologic reporting of acute traumatic brain injuries: A CENTER-TBI Study. J Neurotrauma 2020 Jun 01;37(11):1269-1282 [FREE Full text] [doi: 10.1089/neu.2019.6710] [Medline: 31813313]
29. Schiariti V, Fowler E, Brandenburg JE, Levey E, Mcintyre S, Sukal-Moulton T, et al. A common data language for clinical research studies: the National Institute of Neurological Disorders and Stroke and American Academy for Cerebral Palsy and Developmental Medicine Cerebral Palsy Common Data Elements Version 1.0 recommendations. Dev Med Child Neurol 2018 Oct 15;60(10):976-986 [FREE Full text] [doi: 10.1111/dmcn.13723] [Medline: 29542813]
30. Hunt C, Michalak A, Ouchterlony D, Marshall S, Masanic C, Vaidyanath C, et al. Common Data Elements for Concussion in Tertiary Care: Phase One in Ontario. Can J Neurol Sci 2017 Nov 30;44(6):676-683. [doi: 10.1017/cjn.2017.222] [Medline: 29391082]
31. Mawji A, Li E, Chandna A, Kortz T, Akech S, Wiens MO, et al. Common data elements for predictors of pediatric sepsis: a framework to standardize data collection. PLoS One 2021;16(6):e0253051 [FREE Full text] [doi: 10.1371/journal.pone.0253051] [Medline: 34111209]
32. de Oliveira Manoel AL, van der Jagt M, Amin-Hanjani S, Bambakidis NC, Brophy GM, Bulsara K, Unruptured AneurysmsSAH − CDE Project Investigators. Common data elements for unruptured intracranial aneurysms and aneurysmal subarachnoid hemorrhage: recommendations from the Working Group on Hospital Course and Acute Therapies-Proposal of a Multidisciplinary Research Group. Neurocrit Care 2019 Jun;30(Suppl 1):36-45. [doi: 10.1007/s12028-019-00726-3] [Medline: 31119687]
33. Le Gal G, Carrier M, Castellucci LA, Cuker A, Hansen J, Klok FA, ISTH CDE Task Force. Development and implementation of common data elements for venous thromboembolism research: on behalf of SSC Subcommittee on official Communication from the SSC of the ISTH. J Thromb Haemost 2021 Jan;19(1):297-303 [FREE Full text] [doi: 10.1111/jth.15138] [Medline: 33405381]
34. Gagnon I, Friedman D, Beauchamp MH, Christie B, DeMatteo C, Macartney G, et al. The Canadian Pediatric Mild Traumatic Brain Injury Common Data Elements Project: Harmonizing Outcomes to Increase Understanding of Pediatric Concussion. J Neurotrauma 2018 Aug 15;35(16):1849-1857. [doi: 10.1089/neu.2018.5887] [Medline: 30074870]
35. Massey KA, Magee LA, Dale S, Claydon J, Morris TJ, von Dadelszen P, et al. A current landscape of provincial perinatal data collection in canada. J Obstet and Gynaecol Canada 2009 Mar;31(3):236-246. [doi: 10.1016/s1701-2163(16)34122-6]
36. Lau F, Downing M, Tayler C, Fassbender K, Lesperance M, Barnett J. Toward a population-based approach to end-of-life care surveillance in Canada: initial efforts and lessons. J Palliat Care 2018 Dec 19;29(1):13-21. [doi: 10.1177/082585971302900103]

37. Yang Y, Xu H, Qi B, Niu X, Li M, Zhao D. Stroke screening data modeling based on openEHR and NINDS Stroke CDE. 2020 Presented at: 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); December 16-19, 2020; Seoul, Korea. [doi: 10.1109/BIBM49941.2020.9313127]
38. Biering-Sørensen F, Alai S, Anderson K, Charlifue S, Chen Y, DeVivo M, et al. Common data elements for spinal cord injury clinical research: a National Institute for Neurological Disorders and Stroke project. Spinal Cord 2015 Apr 10;53(4):265-277 [FREE Full text] [doi: 10.1038/sc.2014.246] [Medline: 25665542]
39. Bernstein K. Reporting of drug allergies for use in a national decision support system. Stud Health Technol Inform 2014;205:68-72. [Medline: 25160147]
40. Thurin NH, Pajouheshnia R, Roberto G, Dodd C, Hyeraci G, Bartolini C, et al. From Inception to ConcePTION: Genesis of a Network to Support Better Monitoring and Communication of Medication Safety During Pregnancy and Breastfeeding. Clin Pharmacol Ther 2022 Jan 26;111(1):321-331 [FREE Full text] [doi: 10.1002/cpt.2476] [Medline: 34826340]
41. Holz C, Kessler T, Dugas M, Varghese J. Core Data Elements in Acute Myeloid Leukemia: A Unified Medical Language System-Based Semantic Analysis and Experts' Review. JMIR Med Inform 2019 Aug 12;7(3):e13554 [FREE Full text] [doi: 10.2196/13554] [Medline: 31407666]
42. Hackenberg KAM, Algra A, Al-Shahi Salman R, Frösen J, Hasan D, Juvela S, Unruptured AneurysmsSAH CDE Project Investigators. Definition and prioritization of data elements for cohort studies and clinical trials on patients with unruptured intracranial aneurysms: proposal of a multidisciplinary research group. Neurocrit Care 2019 Jun;30(Suppl 1):87-101. [doi: 10.1007/s12028-019-00729-0] [Medline: 31102238]
43. von Martial S, Brix TJ, Klotz L, Neuhaus P, Berger K, Warnke C, et al. EMR-integrated minimal core dataset for routine health care and multiple research settings: a case study for neuroinflammatory demyelinating diseases. PLoS One 2019;14(10):e0223886 [FREE Full text] [doi: 10.1371/journal.pone.0223886] [Medline: 31613917]
44. Berger A, Rustemeier A, Göbel J, Kadioglu D, Britz V, Schubert K, et al. How to design a registry for undiagnosed patients in the framework of rare disease diagnosis: suggestions on software, data set and coding system. Orphanet J Rare Dis 2021 May 01;16(1):198 [FREE Full text] [doi: 10.1186/s13023-021-01831-3] [Medline: 33933089]
45. Varghese J, Holz C, Neuhaus P, Bernardi M, Boehm A, Ganser A, et al. Key Data Elements in Myeloid Leukemia. Stud Health Technol Inform 2016;228:282-286. [Medline: 27577388]
46. Schaaf J, Chalmers J, Omran H, Pennekamp P, Sitbon O, Wagner T, et al. The Registry Data Warehouse in the European Reference Network for Rare Respiratory Diseases - background, conception and implementation. Stud Health Technol Inform 2021 May 24;278:41-48. [doi: 10.3233/SHTI210049] [Medline: 34042874]
47. Moratilla JM, Alonso-Calvo R, Molina-Vaquero G, Paraiso-Medina S, Perez-Rey D, Maojo V. A data model based on semantically enhanced HL7 RIM for sharing patient data of breast cancer clinical trials. Stud Health Technol Inform 2013;192:971. [Medline: 23920745]
48. Rho MJ, Kim SR, Park SH, Jang KS, Park BJ, Hong JY, et al. Common data model for decision support system of adverse drug reaction to extract knowledge from multi-center database. Inf Technol Manag 2015 Jul 3;17(1):57-66. [doi: 10.1007/s10799-015-0240-6]
49. Cohen JM, Cesta CE, Kjerpeseth L, Leinonen MK, Hálfdánarson Ó, Karlstad Ø, et al. A common data model for harmonization in the Nordic Pregnancy Drug Safety Studies (NorPreSS). Nor J Epidemiol 2021 Aug 16;29(1-2):117-123. [doi: 10.5324/nje.v29i1-2.4053]
50. Krüger A, Lockey D, Kurola J, Di Bartolomeo S, Castrén M, Mikkelsen S, et al. A consensus-based template for documenting and reporting in physician-staffed pre-hospital services. Scand J Trauma Resusc Emerg Med 2011 Nov 23;19(1):71 [FREE Full text] [doi: 10.1186/1757-7241-19-71] [Medline: 22107787]
51. Ringdal KG, Lossius HM, SCANTEM ad hoc group on Scandinavian MTOSTrauma Registry. Feasibility of comparing core data from existing trauma registries in scandinavia. Reaching for a Scandinavian major trauma outcome study (MTOS). Scand J Surg 2007 Jun 24;96(4):325-331. [doi: 10.1177/145749690709600412] [Medline: 18265862]
52. Jaboyedoff M, Rakic M, Bachmann S, Berger C, Diezi M, Fuchs O, et al. SwissPedData: standardising hospital records for the benefit of paediatric research. Swiss Med Wkly 2021 Dec 20;151:w30069 [FREE Full text] [doi: 10.4414/smw.2021.w30069] [Medline: 34964587]
53. Chen S, Hsu C, Huang C. Annotating Taiwan Cancer Registry to caDSR for International Interoperability. In: Zhang Y, editor. Future Communication, Computing, Control and Management. Lecture Notes in Electrical Engineering, vol 141. Berlin, Heidelberg: Springer; 2012:257-263.
54. Spronk P, Begum H, Vishwanath S, Crosbie A, Earnest A, Elder E, et al. Toward International Harmonization of Breast Implant Registries: International Collaboration of Breast Registry Activities Global Common Data Set. Plast Reconstr Surg 2020 Aug;146(2):255-267. [doi: 10.1097/PRS.0000000000006969] [Medline: 32740572]
55. Mowforth O, Khan D, Wong M, Pickering G, Dean L, Magee J, AO Spine RECODE-DCM Steering CommitteeAO Spine RECODE-DCM Consortium. Gathering Global Perspectives to Establish the Research Priorities and Minimum Data Sets for Degenerative Cervical Myelopathy: Sampling Strategy of the First Round Consensus Surveys of AO Spine RECODE-DCM. Global Spine J 2022 Feb;12(1_suppl):8S-18S [FREE Full text] [doi: 10.1177/21925682211047546] [Medline: 34879754]

56. Currie AC, Cahill R, Delaney CP, Faiz OD, Kennedy RH. International expert consensus on endpoints for full-thickness laparoendoscopic colonic excision. Surg Endosc 2016 Apr 27;30(4):1497-1502. [doi: 10.1007/s00464-015-4362-z] [Medline: 26123345]
57. Davies BM, Khan DZ, Mowforth OD, McNair AGK, Gronlund T, Kolias AG, et al. RE-CODE DCM (REsearch Objectives and Common Data Elements for Degenerative Cervical Myelopathy): A Consensus Process to Improve Research Efficiency in DCM, Through Establishment of a Standardized Dataset for Clinical Research and the Definition of the Research Priorities. Global Spine J 2019 May 08;9(1 Suppl):65S-76S [FREE Full text] [doi: 10.1177/2192568219832855] [Medline: 31157148]
58. Cohen D, Sullivan K, McNeil R, Gulf War Illness Common Data Elements Working Group:, Symptoms Assessment Working Group:, McNeil R, Systems Assessment Working Group:, et al. A common language for Gulf War Illness (GWI) research studies: GWI common data elements. Life Sci 2022 Feb 01;290:119818 [FREE Full text] [doi: 10.1016/j.lfs.2021.119818] [Medline: 34352259]
59. Karaa A, Rahman S, Lombès A, Yu-Wai-Man P, Sheikh MK, Alai-Hansen S, Mito Working Group Member Participants:. Common data elements for clinical research in mitochondrial disease: a National Institute for Neurological Disorders and Stroke project. J Inherit Metab Dis 2017 May 16;40(3):403-414 [FREE Full text] [doi: 10.1007/s10545-017-0035-5] [Medline: 28303425]
60. Suarez JI, Sheikh MK, Macdonald RL, Amin-Hanjani S, Brown RD, de Oliveira Manoel AL, Unruptured Intracranial AneurysmsSAH CDE Project Investigators. Common Data Elements for Unruptured Intracranial Aneurysms and Subarachnoid Hemorrhage Clinical Research: A National Institute for Neurological Disorders and Stroke and National Library of Medicine Project. Neurocrit Care 2019 Jun 13;30(Suppl 1):4-19. [doi: 10.1007/s12028-019-00723-6] [Medline: 31087257]
61. Grinspan ZM, Patel AD, Shellhaas RA, Berg AT, Axeen ET, Bolton J, Pediatric Epilepsy Learning Healthcare System. Design and implementation of electronic health record common data elements for pediatric epilepsy: Foundations for a learning health care system. Epilepsia 2021 Jan 24;62(1):198-216 [FREE Full text] [doi: 10.1111/epi.16733] [Medline: 33368200]
62. Ward S, Flori H, Bennett T, Sapru A, Mourani P, Thomas N, et al. Design and Rationale for Common Data Elements for Clinical Research in Pediatric Critical Care Medicine. Pediatr Crit Care Med 2020 Nov;21(11):e1038-e1041 [FREE Full text] [doi: 10.1097/PCC.0000000000002455] [Medline: 32639472]
63. Luciano MG, Batzdorf U, Kula RW, Rocque BG, Maher CO, Heiss J, Chiari I Malformation Common Data Element Working Group. Development of common data elements for use in chiari malformation type I clinical research: an NIH/NINDS project. Neurosurgery 2019 Dec 01;85(6):854-860 [FREE Full text] [doi: 10.1093/neuros/nyy475] [Medline: 30690581]
64. Mayer CS, Williams N, Huser V. Identification of common data elements from pivotal FDA trials. AMIA Annu Symp Proc 2020;2020:813-822 [FREE Full text] [Medline: 33936456]
65. Broglio SP, Kontos AP, Levin H, Schneider K, Wilde EA, Cantu RC, et al. National Institute of Neurological Disorders and Stroke and Department of Defense Sport-Related Concussion Common Data Elements Version 1.0 recommendations. J Neurotrauma 2018 Dec 01;35(23):2776-2783 [FREE Full text] [doi: 10.1089/neu.2018.5643] [Medline: 29717643]
66. Wandner L, Domenichiello A, Beierlein J, Pogorzala L, Aquino G, Siddons A, NIH Pain Consortium Institute and Center Representatives. NIH's Helping to End Addiction Long-term Initiative (NIH HEAL Initiative) clinical pain management common data element program. J Pain 2022 Mar;23(3):370-378 [FREE Full text] [doi: 10.1016/j.jpain.2021.08.005] [Medline: 34508905]
67. Carroll A, Vogel LC, Zebracki K, Noonan VK, Biering-Sørensen F, Mulcahey MJ. Relevance of the international spinal cord injury basic data sets to youth: an Inter-Professional review with recommendations. Spinal Cord 2017 Sep 28;55(9):875-881. [doi: 10.1038/sc.2017.14] [Medline: 28244501]
68. Albino J, Tiwari T, Gansky S, Henshaw M, Barker J, Brega A, Early Childhood Caries Collaborating Centers. The basic research factors questionnaire for studying early childhood caries. BMC Oral Health 2017 May 19;17(1):83 [FREE Full text] [doi: 10.1186/s12903-017-0374-5] [Medline: 28526003]
69. Fisher J, Krisa L, Middleton D, Leiby B, Harrop J, Shah L, et al. Validation of the National Institute of Neurological Disorders and Stroke Spinal Cord Injury MRI Common Data Elements Instrument. AJNR Am J Neuroradiol 2021 Feb 11;42(4):787-793. [doi: 10.3174/ajnr.a7000]
70. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc 2012 Jan 01;19(1):54-60 [FREE Full text] [doi: 10.1136/amiajnl-2011-000376] [Medline: 22037893]
71. Gardner D, Knuth KH, Abato M, Erde SM, White T, DeBellis R, et al. Common data model for neuroscience data and data model exchange. J Am Med Inform Assoc 2001 Jan 01;8(1):17-33 [FREE Full text] [doi: 10.1136/jamia.2001.0080017] [Medline: 11141510]
72. Patel AA, Kajdacsy-Balla A, Berman JJ, Bosland M, Datta MW, Dhir R, et al. The development of common data elements for a multi-institute prostate cancer tissue bank: the Cooperative Prostate Cancer Tissue Resource (CPCTR) experience. BMC Cancer 2005 Aug 21;5(1):108 [FREE Full text] [doi: 10.1186/1471-2407-5-108] [Medline: 16111498]

73. Reisinger SJ, Ryan PB, O'Hara DJ, Powell GE, Painter JL, Pattishall EN, et al. Development and evaluation of a common data model enabling active drug safety surveillance using disparate healthcare databases. J Am Med Inform Assoc 2010 Nov 01;17(6):652-662 [FREE Full text] [doi: 10.1136/jamia.2009.002477] [Medline: 20962127]
74. Loring D, Lowenstein D, Barbaro N, Fureman B, Odenkirchen J, Jacobs M, et al. Common data elements in epilepsy research: development and implementation of the NINDS epilepsy CDE project. Epilepsia 2011 Jun;52(6):1186-1191 [FREE Full text] [doi: 10.1111/j.1528-1167.2011.03018.x] [Medline: 21426327]
75. McCauley SR, Wilde EA, Anderson VA, Bedell G, Beers SR, Campbell TF, Pediatric TBI Outcomes Workgroup. Recommendations for the use of common outcome measures in pediatric traumatic brain injury research. J Neurotrauma 2012 Mar 01;29(4):678-705 [FREE Full text] [doi: 10.1089/neu.2011.1838] [Medline: 21644810]
76. Gerring JP, Wade S. The essential role of psychosocial risk and protective factors in pediatric traumatic brain injury research. J Neurotrauma 2012 Mar 01;29(4):621-628 [FREE Full text] [doi: 10.1089/neu.2011.2234] [Medline: 22091875]
77. Saver JL, Warach S, Janis S, Odenkirchen J, Becker K, Benavente O, et al. Standardizing the Structure of Stroke Clinical and Epidemiologic Research Data. Stroke 2012 Apr;43(4):967-973. [doi: 10.1161/strokeaha.111.634352]
78. Dastgir J, Rutkowski A, Alvarez R, Cossette S, Yan K, Hoffmann R, et al. Common Data Elements for Muscle Biopsy Reporting. Arch Pathol Lab Med 2016 Jan;140(1):51-65 [FREE Full text] [doi: 10.5858/arpa.2014-0453-OA] [Medline: 26132600]
79. Perrone RD, Neville J, Chapman AB, Gitomer BY, Miskulin DC, Torres VE, et al. Therapeutic area data standards for autosomal dominant polycystic kidney disease: a report from the Polycystic Kidney Disease Outcomes Consortium (PKDOC). Am J Kidney Dis 2015 Oct;66(4):583-590. [doi: 10.1053/j.ajkd.2015.04.044] [Medline: 26088508]
80. Yue JK, Vassar MJ, Lingsma HF, Cooper SR, Okonkwo DO, Valadka AB, TRACK-TBI Investigators. Transforming research and clinical knowledge in traumatic brain injury pilot: multicenter implementation of the common data elements for traumatic brain injury. J Neurotrauma 2013 Nov 15;30(22):1831-1844 [FREE Full text] [doi: 10.1089/neu.2013.2970] [Medline: 23815563]
81. Hicks K, Tcheng J, Bozkurt B, Chaitman B, Cutlip D, Farb A, American College of Cardiology, American Heart Association. 2014 ACC/AHA Key Data Elements and Definitions for Cardiovascular Endpoint Events in Clinical Trials: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Data Standards (Writing Committee to Develop Cardiovascular Endpoints Data Standards). Circulation 2015 Jul 28;132(4):302-361 [FREE Full text] [doi: 10.1161/CIR.0000000000000156] [Medline: 25547519]
82. Moore SM, Schiffman R, Waldrop-Valverde D, Redeker NS, McCloskey DJ, Kim MT, et al. Recommendations of common data elements to advance the science of self-management of chronic conditions. J Nurs Scholarsh 2016 Sep;48(5):437-447 [FREE Full text] [doi: 10.1111/jnu.12233] [Medline: 27486851]
83. Rubinstein YR, McInnes P. NIH/NCATS/GRDR® Common Data Elements: a leading force for standardized data collection. Contemp Clin Trials 2015 May;42:78-80 [FREE Full text] [doi: 10.1016/j.cct.2015.03.003] [Medline: 25797358]
84. De Vito Dabbs A, Myers B, Mc Curry K, Dunbar-Jacob J, Hawkins R, Begey A, et al. User-centered design and interactive health technologies for patients. Comput Inform Nurs 2009;27(3):175-183 [FREE Full text] [doi: 10.1097/NCN.0b013e31819f7c7c] [Medline: 19411947]

Abbreviations

CDE: common data element
CDM: common data model
MeSH: Medical Subject Headings
OMOP: Observational Medical Outcomes Partnership
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
SATURN: Smartes Arztportal für Betroffene mit unklarer Erkrankung (Smart physicians' platform for patients with unclear diseases)

Edited by C Lovis; submitted 16.12.22; peer-reviewed by A Lamer, X Ma; comments to author 31.01.23; revised version received 09.03.23; accepted 08.06.23; published 03.08.23.

Please cite as:
Ahmadi N, Zoch M, Kelbert P, Noll R, Schaaf J, Wolfien M, Sedlmayr M
Methods Used in the Development of Common Data Models for Health Data: Scoping Review
JMIR Med Inform 2023;11:e45116
URL: https://medinform.jmir.org/2023/1/e45116
doi:10.2196/45116
PMID:37535410

Review

Designing Interoperable Health Care Services Based on Fast Healthcare Interoperability Resources: Literature Review

Jingwen Nan[1], MA; Li-Qun Xu[1], PhD

Health IT Research, China Mobile (Chengdu) Industrial Research Institute, Chengdu, China

Corresponding Author:
Li-Qun Xu, PhD
Health IT Research
China Mobile (Chengdu) Industrial Research Institute
Unit 2, Block C1, AI Innovation Center
Hele Second Street, Gaoxin District
Chengdu, 610213
China
Phone: 86 (28) 60103585
Email: xuliqun@chinamobile.com

Abstract

Background: With the advent of the digital economy and the aging population, the demand for diversified health care services and innovative care delivery models has been overwhelming. This trend has accelerated the urgency to implement effective and efficient data exchange and service interoperability, which underpins coordinated care services among tiered health care institutions, improves the quality of oversight of regulators, and provides vast and comprehensive data collection to support clinical medicine and health economics research, thus improving the overall service quality and patient satisfaction. To meet this demand and facilitate the interoperability of IT systems of stakeholders, after years of preparation, Health Level 7 formally introduced, in 2014, the Fast Healthcare Interoperability Resources (FHIR) standard. It has since continued to evolve. FHIR depends on the Implementation Guide (IG) to ensure feasibility and consistency while developing an interoperable health care service. The IG defines rules with associated documentation on how FHIR resources are used to tackle a particular problem. However, a gap remains between IGs and the process of building actual services because IGs are rules without specifying concrete methods, procedures, or tools. Thus, stakeholders may feel it nontrivial to participate in the ecosystem, giving rise to the need for a more actionable practice guideline (PG) for promoting FHIR's fast adoption.

Objective: This study aimed to propose a general FHIR PG to facilitate stakeholders in the health care ecosystem to understand FHIR and quickly develop interoperable health care services.

Methods: We selected a collection of FHIR-related papers about the latest studies or use cases on designing and building FHIR-based interoperable health care services and tagged each use case as belonging to 1 of the 3 dominant innovation feature groups that are also associated with practice stages, that is, data standardization, data management, and data integration. Next, we reviewed each group's detailed process and key techniques to build respective care services and collate a complete FHIR PG. Finally, as an example, we arbitrarily selected a use case outside the scope of the reviewed papers and mapped it back to the FHIR PG to demonstrate the effectiveness and generalizability of the PG.

Results: The FHIR PG includes 2 core elements: one is a practice design that defines the responsibilities of stakeholders and outlines the complete procedure from data to services, and the other is a development architecture for practice design, which lists the available tools for each practice step and provides direct and actionable recommendations.

Conclusions: The FHIR PG can bridge the gap between IGs and the process of building actual services by proposing actionable methods, procedures, and tools. It assists stakeholders in identifying participants' roles, managing the scope of responsibilities, and developing relevant modules, thus helping promote FHIR-based interoperable health care services.

(JMIR Med Inform 2023;11:e44842) doi:10.2196/44842

KEYWORDS

Health level 7 Fast Healthcare Interoperability Resources; HL7 FHIR; interoperability; literature review; practice guideline; mobile phone

Introduction

Background

The development and innovation of health care service models have accelerated the demand for data exchange and service interoperability. In the United States, the Health Information Technology for Economic and Clinical Health Act took effect in 2009, specifying health IT–based systems as an integrated part of the country's health care reform. It has spurred the electronic health record (EHR) adoption rate through reward and punishment measures [1]. In addition, the US Department of Health and Human Services established a specific agency, the Office of the National Coordinator for Health Information Technology, to accelerate the implementation of advanced medical IT standards, promote the exchange of electronic health care information, and improve the quality of health care services throughout the country. In Canada, the federal government funded an independent, not-for-profit organization called Canada Health Infoway, tasked with accelerating the adoption of digital health solutions, such as EHR, across the country. The government has set a 10-year implementation strategy for EHR in cooperation with the Canadian Institute for Health Information [2]. Japan has made great efforts to develop remote health care technology and has established a communication system among regional institutions by implementing electronic medical records (EMRs) in the form of an app or software as a service [3]. In China's state health system, major public hospitals administered by national, provincial, and local health authorities are the pioneers in reforms. Over the years, the government has issued a series of policies promoting coordinated care among health care institutions at different levels of the health system [4,5], together with many qualitative or quantitative assessment criteria that guide the establishment of high-standard EMR system, regional information interoperability, and intelligent service and management in hospitals. In summary, the demand for tiered and coordinated care delivery among health care institutions worldwide is increasing rapidly, and the requirement for health care data exchange continues unabated.

The enhancement of interoperability is required by transforming health care service models and tackling the challenges of societal problems. According to a United Nations report [6], the share of the population aged ≥65 years is expected to increase from 9.3% in 2020 to approximately 16% in 2050. The rapid aging of the population unavoidably increases the burden of chronic disease care, bringing about the requirements for people-centered and continuous care delivery built on the foundation of a robust primary health care system. Therefore, it is necessary to enhance health IT system interoperability to bridge the gap between uneven health care resource distribution, remove the barrier of isolated data islands, and comprehensively improve the quality of health care services.

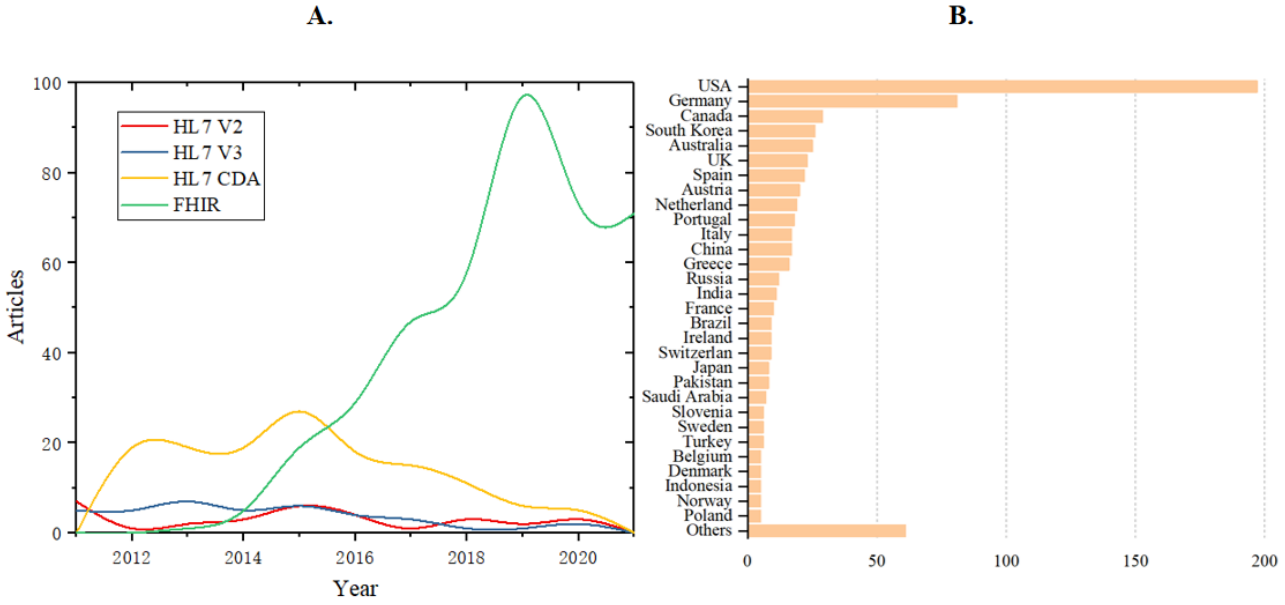
Health Level 7 Fast Healthcare Interoperability Resources

Health Level 7 (HL7), founded in 1987, is a not-for-profit, standards-developing organization dedicated to providing a comprehensive framework and related standards for the exchange, integration, sharing, and retrieval of electronic health information that supports clinical practice and the management, delivery, and evaluation of health services. It has successively released many standards, including HL7 version 2, HL7 version 3, and Clinical Document Architecture (CDA). However, with the constant evolution of the internet and the thriving of the application programming interface (API) economy, digital services or assets of health organizations tend to be exposed even more widely in the form of APIs. In this context, HL7 formally introduced Fast Healthcare Interoperability Resources (FHIR) in 2014, highlighting the core concept of resources, and thus, creating a new era for health care service interoperability. A resource is the smallest exchangeable logical unit in FHIR. Resources are independent of each other but can be linked or assembled through specific rules to meet diverse service requirements. FHIR combines web standards to support resource operations through RESTful API in XML or JavaScript Object Notation format. Compared with other alternative standards, FHIR has more advantages and potential, such as comprehensive coverage of data definitions, substantial flexibility of data exchange, explicit semantics, and many available open-source tools, among others. Therefore, it has attracted constant and favorable attention from health care stakeholders since its first release, as shown in Figure 1.

We investigated the literature from the Web of Science and plotted 2 statistical charts in Figure 1. Figure 1A shows the promotion trends of different health data standards. By using the search term "HL7 v2," "HL7 v3," "HL7 CDA," and "FHIR," we identified the corresponding papers in the Web of Science database from 2010 to 2022. The results show that the attention paid to FHIR has increased rapidly within a short time, far exceeding the HL7 version 2, HL7 version 3, and CDA standards. Figure 1B compares FHIR-relevant literature among different countries. We used the search term "FHIR" to find the corresponding papers in the Web of Science database from 2014 to 2022. By reading each paper's abstract and the corresponding author's information, we identified the country to which the work belongs. Countries that record <5 papers fall into the "others" category. The chart shows that the United States, Germany, and Canada were the top 3 countries that published the most studies on FHIR, accounting for 28.39% (197/694), 11.67% (81/694), and 4.18% (29/694), respectively.

In addition to the dissemination activities of enthusiastic researchers and pioneering health IT ecosystem players, national health policy makers also play a pivotal role in FHIR adoption, as evidenced by the actions in the United Kingdom, United States, and Canada [7]. Overall, FHIR has gradually gained worldwide recognition and acceptance, and it has the most potential for future large-scale promotion in the health care ecosystem.

Figure 1. Works of literature that focus on health data standards. (A) The attention to Fast Healthcare Interoperability Resources (FHIR) has risen rapidly within a short time of its first release, far exceeding HL7 version 2, HL7 version 3, and Clinical Document Architecture (CDA) standards. (B) The United States, Germany, and Canada are the top 3 countries that published the most literature on FHIR. HL: Health Level.



Objectives

Owing to the growing popularity of FHIR, some academic researchers have authored review papers from their perspectives in the last few years. Ayaz et al [8] searched for FHIR-related papers published between 2012 and 2019 in 6 databases (ACM, IEEE, Springer, Google Scholar, PubMed, and ScienceDirect) and selected 80 papers for review. They found that FHIR is identical in supporting intelligent technologies, such as smartphones, tablets, mobile health apps, smartwatches, and fitness trackers, which could solve numerous health care problems that were impossible for the previous standards. Lehne et al [9] searched for FHIR-related papers in 2 databases (Web of Science and PubMed) up to 2019 and selected 131 papers for review. The statistical results revealed that data model–related topics mainly focusing on constructing profiles to implement FHIR in specific scenarios were the most attractive direction. At the same time, analytics-related topics concerning data analysis, modeling, machine learning, and more were less attractive because most FHIR projects were still in the initial development phase, dealing with implementation and data definitions rather than large-scale data analysis. Barker and Johnson[10] surveyed 734 apps released up to December 2020 in 5 digital health care application libraries (hosted by Cerner, Epic, Allscripts, Athenahealth, and Substitutable Medical Applications Reusable Technologies [SMART]) and measured their support for FHIR. They found that the number of apps that support the FHIR standard had increased from 19% in 2019 to 22% in 2020.

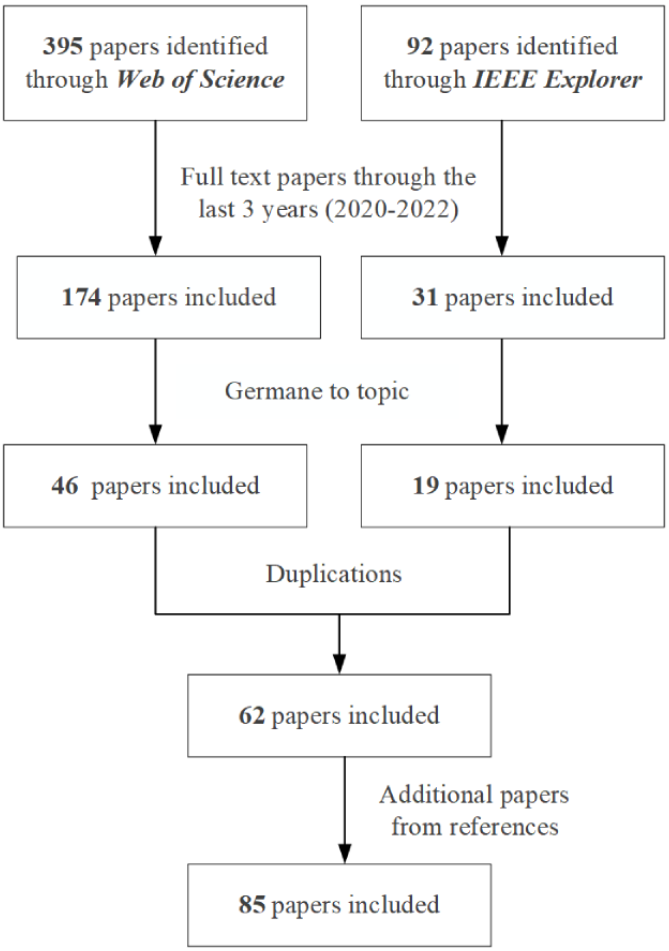
However, to our knowledge, there is a lack of systematic reviews that focus on the FHIR practice. A gap remains between the FHIR Implementation Guide (IG) and building actual services because IGs are rules specifying no methods, procedures, or tools. Thus, stakeholders may feel it nontrivial to participate in the ecosystem, giving rise to the need for a more actionable practice guideline (PG) for promoting FHIR's fast adoption. Therefore, this study proposed a general FHIR PG to facilitate stakeholders in the health care ecosystem to understand FHIR and quickly develop interoperable health care services.

Methods

Article Selection

Figure 2 presents the paper selection flowchart used in this review. Initially, we identified a total of 487 papers in the Web of Science and IEEE databases by using the search term "FHIR" or "Fast Healthcare Interoperability Resources." The time range of publications was set from January 1, 2020, to July 1, 2022, and we finalized 205 articles. After excluding those that merely mentioned the term FHIR but did not elaborate on it, 65 articles were retained. A check of duplications from this batch removed a further 3 articles. Finally, from the references of the remaining 62 articles, we found an additional 23 relevant articles, ending up with a total of 85 articles as the research materials of this study.

Figure 2. Flowchart of paper selection.



Analysis Process

By carefully analyzing and collating the recent studies on the design of FHIR-based interoperable health care services, we derived the details of the FHIR PG.

We selected 85 FHIR-related articles and found that building FHIR-based health care services contains typically 3 stages, that is, data standardization, data management, and data integration. Each stage may use different practice methods, depending on the targeted scenarios and types of services.

The way to categorize these 85 articles is as follows: if an article's main innovation feature focused on 1 of the 3 stages, we assigned it to the corresponding group. Specifically, we assigned those articles emphasizing the design process of FHIR profiles or proposing methods for migrating data from specific clinical data models (CDMs) to FHIR to the data standardization group, articles discussing the management of RESTful APIs to the data management group, and articles presenting approaches for integrating data with specific apps or platforms to the data integration group.

After categorizing the articles, we reviewed the key techniques used by each group to build their respective health care services. We compiled a general FHIR PG through this review. The workflow of the FHIR PG was derived by linking the stages, each consisting of multiple steps. It is important to note that alternative solutions might be identified for certain steps in the workflow based on different conditions. In addition, we leveraged the collective experience of our team working on health care IT projects to further refine and optimize the FHIR PG.

Finally, as an example, we arbitrarily selected a use case outside the scope of the reviewed articles and mapped it back to the FHIR PG to demonstrate the effectiveness and generalizability of the PG.

Results

Article Classification

Data Standardization

Data standardization typically involves two main steps: (1) defining profiles based on the data exchange requirements of interoperable services and (2) filling these profiles with the corresponding exchange data.

The base FHIR specification provides foundational resources applicable to various health care contexts. However, health care services often exhibit significant variability across different jurisdictions. Therefore, the base FHIR specification typically requires further adaptation, known as profile definition, to suit specific application contexts. Profile definition mainly encompasses three aspects: (1) rules about which resource elements to use and what additional elements to add to the base specification, (2) rules about which terminologies to use in

particular elements, and (3) the restricted value range and cardinality of the elements.

Table 1 lists the typical profile definitions and the corresponding FHIR foundational resources discussed in the reviewed articles. As shown, these articles cover a wide range of categories, including genomics [10-14], imaging [15-17], cancer [18-20], diabetes [21,22], COVID-19 [23,24], infections [25], electrocardiography [26], screening [27], and allergy [28].

There are typically 2 approaches to filling the profiles with exchange data. One is redesigning the database to align with the FHIR resource structure, and the other is mapping data from an existing CDM-based legacy system to the FHIR-based system. Table 2 lists relevant articles discussing the latter approach. These articles could roughly fall into 7 groups based on the types of source CDMs. The groups include informatics for integrating biology and the bedside [29,30], Observational Medical Outcomes Partnership (OMOP) [31,32], OpenEHR [33,34], HL7 version 2 [35], variant call format [36], free text or arbitrary proprietary data [37,38], and multisource [39-42]. Multisource refers to cases where multiple CDMs are involved. For example, the study by Lenert et al [40] focused on transforming data from the OMOP and Patient-Centered Outcomes Research Network to FHIR. The study by Pfaff et al [39] aimed to transform data from informatics for integrating biology and the bedside, OMOP, and Patient-Centered Outcomes Research Network to FHIR. The study by Prud'hommeaux et al [41] compared 3 methods for transforming data from various source CDMs into FHIR. The study by Kiourtis et al [42] proposed a resource description framework transformation toolkit to combine FHIR and non-FHIR data.

The studies in Table 2 indicate that the transformation from a specific CDM type to FHIR typically involves a 2-step mapping process: model mapping and element mapping. Model mapping establishes a relationship between the original data model and the FHIR resource. Element mapping comprises 2 parts, key mapping and value mapping, which define how to map the data fields from the source CDM to the corresponding fields in the FHIR resources. The mapping rules observe the consensus-mapping relationships established by domain experts. These experts analyzed the semantic and structural differences between the source CDMs and FHIR and determined the appropriate mappings to ensure accurate and meaningful data transformation. Although current data transformation approaches intend to support specific source data and target FHIR resource types, it is worth noting that ongoing research and advancements in domain-based applied artificial intelligence, including natural language processing and deep learning, hold great potential for developing more generalized data transformation algorithms.

As highlighted in previous studies, the granularity of data plays a crucial role in data standardization. When the granularity of the source data is finer than that of the target data, there is potential for information loss during the transformation process: the severity of information loss increases with the extent of the granularity gap.

Table 1. Profile definitions from the reviewed articles.

Theme and study, year	Involved Fast Healthcare Interoperability Resources
Genomics	
Murugan et al [10], 2021	DiagnosticReport, Specimen, ServiceRequest, Observation, and Task
Seong et al [11], 2021	MolecularSequence
Alterovitz et al [12], 2020	DiagnosticReport, ServiceRequest, and Observation
Klopfenstein et al [13], 2021	Questionnaire and Document
Khalifa et al [14], 2021	Patient, PractitionerRole, Organization, Specimen, ServiceRequest, Media, RiskAssessment, Task, MedicationRequest, CarePlan, DeviceRequest, NutritionOrder, SupplyRequest, and RequestGroup
Imaging	
Kohli et al [15], 2018	Patient, DiagnosticReport, ImagingStudy, AllergyIntolerance, Condition, MedicationOrder, Specimen, Organization, Practitioner, and Medication
Madrigal and Le [16], 2021	Media
Boufahja et al [17], 2021	Observation
Cancer	
Zong et al [18], 2021	Observation and DiagnosticReport
Gonzalez-Castro et al [19], 2021	Observation, Device, FamilyMemberHistory, AllergyIntolerance, Condition, Patient, MedicationStatement, Encounter, Questionnaire, QuestionnaireResponse, and Procedure
Zong et al [20], 2020	QuestionnaireResponse
Diabetes	
Ludmann et al [21], 2020	Observation
Glachs et al [22], 2020	Procedure, ProcedureRequest, Communication, Appointment, Observation, Condition, CommunicationRequest, Device, Encounter, Composition, Goal, Order, OrderResponse, MedicationAdministration, MedicationOrder, Organization, Patient, Practitioner, RiskAssessment, QuestionnaireResponse, Basic, and Parameters
COVID-19	
Bauer et al [23], 2021	Questionnaire
Sass et al [24], 2020	Procedure, Observation, Condition, DiagnosticReport, Procedure, Consent, Immunization, MedicationStatement
Infections	
Shivers et al [25], 2021	Consent, Coverage, DeviceUseStatement, Encounter, HealthcareService, Medication, MedicationAdministration, MedicationStatement, Observation, Patient, Practitioner, Procedure, ServiceRequest, and Specimen
Electrocardiogram	
Benhamida et al [26], 2020	Observation
Neonatal screening	
Bathelt et al [27], 2020	Patient, ServiceRequest, DiagnosticReport, Contract, Organization, and Practitioner
Allergy	
Lenivtceva and Kopanitsa [28], 2021	AllergyIntolerance

Table 2. Data migration from the existing clinical data model to Fast Healthcare Interoperability Resources.

Study, year	Clinical data model of the source
Boussadi and Zapletal [29], 2017; Wagholikar et al [30], 2017	Informatics for integrating biology and the bedside
Jiang et al [31], 2017; Fischer et al [32], 2020	Observational Medical Outcomes Partnership
Ladas et al [33], 2022; Fette et al [34], 2020	OpenEHR
Xiao et al [35], 2021	HL7[a] version 2
Dolin et al [36], 2021	Variant call format
Peterson et al [37], 2020; Wang et al [38], 2020	Free text or arbitrary proprietary
Lenert et al [40], 2021; Pfaff et al [39], 2019; Prud'hommeaux et al [41], 2021; Kiourtis et al [42], 2020	Multisource

[a]HL7: Health Level 7.

Data Management

Data management includes data storage and data exposure. Although FHIR defines 5 approaches for data exposure, including RESTful API, messaging, documents, services, and persistent store, recent articles predominantly chose to expose data in the form of APIs because of the rapid growth of the APIs economy. There are typically 2 methods for data management: developing a customized FHIR warehouse to store and manage FHIR data or selecting a mature third-party warehouse to handle the task.

Table 3 shows various data management choices and their corresponding targets. It reveals that developing a customized FHIR warehouse to maintain FHIR data often requires meeting some special service requirements. For instance, the customized FHIR warehouse developed by Demurjian et al [43] aimed to enable sensitivity and multilevel security controls. The one developed by Chatterjee et al [44] and Saripalle et al [45] served to integrate with specific terminology. The one developed by Ruminski et al [46], Saripalle [47], and Yu et al [48] intended to support multiple Internet of Things protocols. Finally, the one discussed in the studies by Khvastova et al [49], Dridi et al [50], Lee et al [51], Tanaka and Yamamoto [52], Cheng et al [53], Semenov et al [54], and Gruendner et al [55] was used to support data preprocess plug-ins.

On the other hand, several mature third-party platforms are available for managing FHIR data. In 2018, a total of 6 technology giants, including Amazon, Microsoft, Google, IBM, Oracle, and Salesforce, jointly announced that they would be committed to removing the barriers to adopting health care interoperability technologies, particularly those enabled through the cloud [56]. All these companies have launched FHIR data management platforms, providing FHIR data APIs for resource operations. Users of these platforms can store their data as FHIR resources and use the data APIs offered by the cloud platform for service development. For instance, the studies by Shi et al [57], Zampognaro et al [58], Ploner and Prokosch [59], and Kamel and Nagy [60] chose cloud warehouses, and the study by Mandl et al [61] chose an on-premises warehouse to rapidly deploy an FHIR development environment.

The abovementioned analysis highlights that choosing between proprietary and third-party warehouses involves trade-off considerations. Maintaining FHIR data through a proprietary warehouse offers 2 advantages: better privacy and greater flexibility for functional expansion. However, developing a proprietary warehouse requires extensive knowledge of FHIR standards and software development skills, resulting in higher costs. On the other hand, relying on third-party platforms offers the advantages of lower cost and higher implementation efficiency. However, storing sensitive data in a third-party warehouse, with the service provider not being the data owner, raises security and privacy concerns.

Table 3. Fast Healthcare Interoperability Resources (FHIR) data management methods and their corresponding targets.

Method and study, year	Target
Develop FHIR warehouse	
Demurjian et al [43], 2020	Support lattice-based access control
Chatterjee et al [44], 2022; Saripalle et al [45], 2020	Integrate with specific terminologies
Ruminski et al [46], 2016; Saripalle [47], 2019; Yu et al [48], 2021	Support multiple IoT[a] protocols
Khvastova et al [49], 2020; Dridi et al [50], 2020; Lee et al [51], 2020; Tanaka and Yamamoto [52], 2020; Cheng et al [53], 2021; Semenov et al [54], 2019; Gruendner et al [55], 2021	Support data preprocess plug-ins
Use third-party FHIR warehouse	
Shi et al [57], 2021; Zampognaro et al [58], 2021 Ploner and Prokosch [59], 2020; Kamel and Nagy [60], 2018	Rapidly deploy a development environment through a cloud FHIR warehouse
Mandl et al [61], 2020	Rapidly deploy a development environment through an on-premises FHIR warehouse

[a]IoT: Internet of Things.

Data Integration

Data integration plays a vital role in health care across various domains, including service delivery, public health management, and clinical medicine or health care economics research, enabling better decision-making and improving overall health care outcomes. In service delivery, data integration is crucial for coordinating multiple IT systems, including the hospital information system (HIS), laboratory information system, picture archiving and communication system, EMR, and EHR. In public health, local governments need to collect health-related data within their jurisdictions to monitor regional health status and effectively address public health issues. In clinical medicine or health care economics research, it is essential to obtain data from diverse domains to conduct comprehensive studies and analyses.

There are 2 typical modes of FHIR data integration, as listed in Table 4.

The first mode of data integration is using an integrated service platform (ISPf). The ISPf is an orchestrating platform offering a series of API management functions such as API registration, API calling authorization, and API routing forward. Organizations wishing to exchange data through the ISPf must register their APIs on the platform. Other organizations can search for the appropriate APIs on the ISPf and make API calls. The ISPf performs API calling authorization to verify the calling rights and then routes the API calls to the respective organization to which the API belongs. This process facilitates data exchange among multiple organizations [62-75]. An example of this mode is the efficient transfer of medical records when a patient referral occurs.

The second mode of data integration is by way of interoperable apps. Different architectures can be selected for different application scenarios. In the case of apps with specific functions, such as statistics and analysis, SMART on FHIR would be a more efficient option [76-85]. In the case of apps with customized functions, such as supporting microservice architecture or blockchain architecture, customized architecture apps would be a more suitable option [86-94].

Table 4. Fast Healthcare Interoperability Resources data integration modes and their corresponding application scenarios.

Interoperable modes and study, year	Applied scenarios
Integrated service platform	
Nan et al [62], 2021; Taechoyotin et al [63], 2021; Maxi and Morocho [64], 2022; Rosenau et al [65], 2022; Corici et al [66], 2020; Papaioannou et al [67], 2021; Hidayat and Hermanto [68], 2020; Sloane et al [69], 2021; Mukhiya and Lamo [70], 2021; Gruendner et al [71], 2022; Gruendner et al [72], 2020; Park et al [73], 2022; Ziminski et al [74], 2021; De et al [75], 2021	Control exchange data through APIs[a] for service coordination among multiple organizations.
App	
Suraj et al [76], 2022; Michaels et al [77], 2021; Curran et al [78], 2020; Thayer et al [79], 2021; Karhade et al [80], 2021; Wesley et al [81], 2021; Burkhardt et al [82], 2021; Hoffman et al [83], 2017; Stoldt and Weber [84], 2020; Stoldt and Weber [85], 2021	Substitutable Medical Applications and Reusable Technologies app: apps with specific functions, such as statistics and analysis.
Alamri et al [86], 2021; George and Chacko [87], 2022; Gulden et al [88], 2021; Chaves et al [89], 2021; Bae and Yi [90], 2022; Bettoni et al [91], 2021; Weber et al [92], 2020; Sfat et al [93], 2021; Mohammed et al [94], 2021	Other architecture app: apps with customized functions, such as supporting microservice and blockchain architecture.

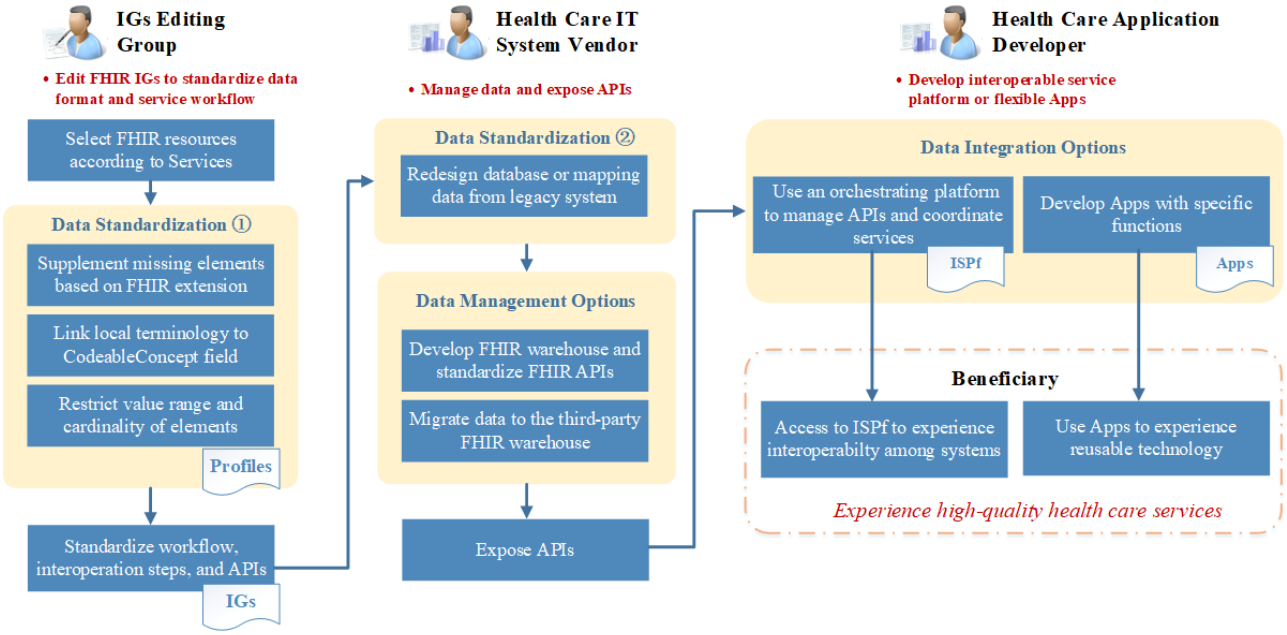
[a]API: application programming interface.

FHIR PG Design

Practice Design

We present an FHIR practice design in Figure 3, which defines the responsibilities of stakeholders and outlines the complete practice process from data to services.

Figure 3. The general Fast Healthcare Interoperability Resources (FHIR) practice guideline—practice design. API: application programming interface; IG: Implementation Guide; ISPf: integrated service platform.



IGs Editing Group

The first stakeholder involved in the process is the IGs editing group, usually coordinated by a government agency or an institution with significant influence in the ecosystem. The primary responsibility of this group is to define the data and service models and release the IGs. The detailed processes are as follows. First, select necessary FHIR resources based on the service requirements. Second, for specific requirements beyond the scope of the original FHIR resources, the group needs to customize resource structure by FHIR profile. Profile generally involves 3 aspects: extending the data field by FHIR extension, linking the local CodeSystem to the CodeableConcept field of FHIR resources, and restricting the cardinality and ValueSet of FHIR foundational resource. The customized resources created by the profile enable better alignment with the data requirements in various scenarios. After completing the data unification task, the IGs editing group moves on to the unification of services workflow, which involves specifying the implementation steps in the workflow and standardizing the corresponding APIs. Ultimately, the abovementioned data and workflow specifications are integrated to form the comprehensive FHIR IGs that health care IT system vendors can adopt.

Health Care IT System Vendor

The second type of stakeholder is the health care IT system vendor, responsible for developing and maintaining systems, such as the HIS, laboratory information system, and picture archiving and communication system. First, the vendor must implement the IGs published by the IGs editing group, which involves standardizing data by redesigning the database according to the FHIR resource structure or mapping data from existing CDM-based legacy systems to FHIR-based systems. Second, with RESTful APIs, the vendor has 2 options for data exposure: either maintaining the FHIR data and APIs themselves or selecting a mature third-party platform. FHIR APIs must be exposed to support resource-level operations regardless of the chosen option.

It is worth pointing out that in terms of data exposure, FHIR defines 5 different approaches, and each data exposure approach has a different data integration method; it would be a lengthy discussion if all approaches are considered. To make FHIR PG more compatible with current technology stacks, we chose to focus on RESTful API rather than on other approaches in this study.

Health Care Application Developer

The third stakeholder involved in this process is the health care application developer, responsible for developing interoperable services using open FHIR APIs. As described in the *Data Integration* section, there are 2 typical modes. The first is to develop an ISPf, that is, an orchestrating platform, for service interoperability. The ISPf manages open APIs registered by each organization and enforces access specifications such as IGs, profiles, and workflows. Any IT systems accessing the ISPf and exchanging data must comply with these specifications. When an IT system needs to access multiple ISPfs, it must support multiple specifications. In such cases, the IT system can deploy an adapter above its native database to comply with various specifications. When the IT system acts as a producer, it reads the corresponding specifications from the adapter to expose the data. When it acts as a consumer, it reads the corresponding specifications from the adapter to parse data. The second mode is to develop specific apps that cater to specific requirements. For example, an app built with SMART on FHIR

architecture supports a flexible and switchable application ecosystem.

Beneficiary

Beneficiaries such as hospitals, patients, public health institutions, and research institutions can benefit from high-quality FHIR-based health care services. For instance, if there is a need to exchange data through APIs to facilitate service coordination among multiple organizations, they can easily access the ISPf to fulfill this objective. Alternatively, they can choose a suitable app from the application gallery that caters to their needs and functions.

The Development Architecture for the Practice Design

Overview

We presented a 3-stage development architecture for the practice design, as shown in Figure 4. In addition, we compiled a list of commonly used tools in Table 5 to support the development process.

Figure 4. The general Fast Healthcare Interoperability Resources (FHIR) practice guideline—the development architecture for the practice design. IG: Implementation Guide; ISPf: integrated service platform; SMART: Substitutable Medical Applications Reusable Technologies.

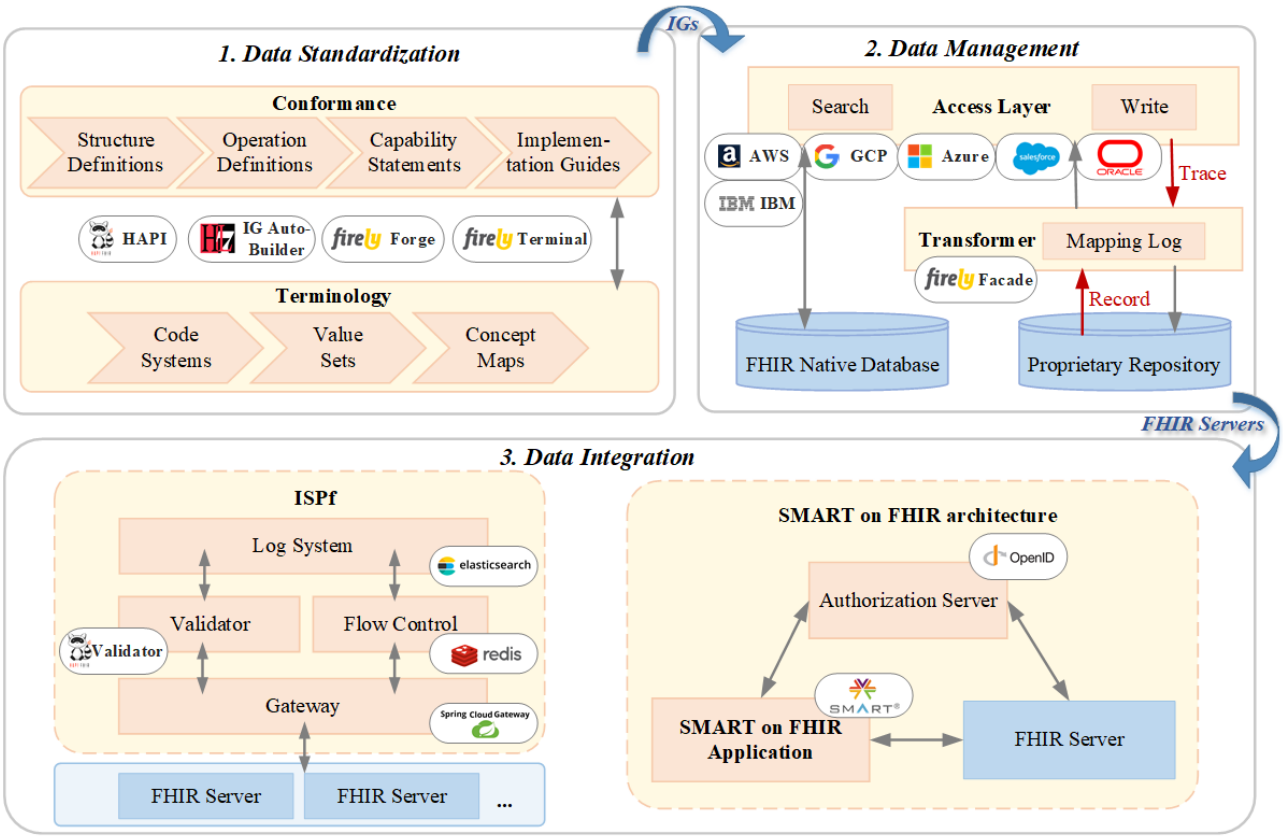


Table 5. A list of commonly used tools.

Tool and description		Availability
Data standardization		
HAPI[a] FHIR[b]	This tool provides Java API[c] for HL7[d] FHIR clients and servers	[95]
IG[e] Auto-Builder	An IG publishing tool that makes your IGs to be visible on the internet [96]	[97]
Firely Forge	The official FHIR tool for managing FHIR profiles	[98]
Firely Terminal	A cross-platform command line tool with a range of commands for working with FHIR resources and installing and publishing FHIR packages	[99]
Data management		
Firely Facade	A special type of plug-in that registers services to access the existing data repository. It speaks FHIR in the front-end and talks directly to native data in the back-end	[100]
FHIR Works on AWS[f]	A framework used to deploy an FHIR server on AWS	[101]
FHIR server for Azure	An open-source implementation of FHIR specification designed for the Microsoft cloud	[102]
GCP[g] Healthcare API	A cloud application that accelerates health care solution development with fully managed, enterprise-scale HL7 FHIR, HL7 version 2, and DICOM[h] APIs	[103]
IBM FHIR server	An open-source Java solution that supports the processing, validation, and storage of health care data according to the HL7 FHIR specification	[104]
Oracle Healthcare Data Repository	The foundation of a health care information exchange platform that makes health care data more useful by supporting the integration and operation of a full spectrum of health care applications	[105]
Health Cloud	A tool that combines clinical and nonclinical customer data to drive efficiencies in health	[106]
Data integration		
Spring Cloud Gateway	This tool provides an API Gateway built on top of the Spring Ecosystem	[107]
Redis	This tool provides access to mutable data structures via a set of commands sent using a server-client model with TCP[i] sockets and a simple protocol	[108]
Validator	The HAPI FHIR Validator API is a simple REST[j] API to validate the structure and content of an FHIR object	[109]
Elasticsearch	A distributed, RESTful search and analytics engine is at the heart of the Elastic Stack	[110]
OpenID	An open standard and decentralized authentication protocol promoted by the nonprofit OpenID Foundation	[111]
OAuth	An open protocol to allow secure authorization in a simple and standard method from web, mobile, and desktop applications	[112]
SMART[k]	Define a workflow that an application can use to securely request access to data and then receive and use that data	[113]

[a]HAPI: Health Level 7 application programming interface.

[b]FHIR: Fast Healthcare Interoperability Resources.

[c]API: application programming interface.

[d]HL7: Health Level 7.

[e]IG: Implementation Guide.

[f]AWS: Amazon Web Services.

[g]GCP: Google Cloud Platform.

[h]DICOM: Digital Imaging and Communications in Medicine.

[i]TCP: transmission control protocol.

[j]REST: representational state transfer.

[k]SMART: Substitutable Medical Applications Reusable Technologies.

Data Standardization

In the data standardization development stage, several components are defined to ensure the consistent use of codes within a specific context. The terminology system comprises essential resources such as CodeSystem, ValueSet, and ConceptMaps. These resources establish a framework for determining which codes can be used. Furthermore, the

conformance system includes resources such as StructureDefinition, OperationDefinition, CapabilityStatement, and ImplementationGuide. These resources are crucial in creating profiles and IGs that adhere to a specific exchange framework. As mentioned in the *Data Standardization* section, the granularity of the data plays a crucial role in information loss. Pfaff et al [39] pointed out that information loss can be avoided by defining custom values or extensions during the data standardization stage. By incorporating custom values or extensions defined in this stage, it is possible to capture and preserve the finer-grained information that is likely to be lost during the transformation process.

During this process, developers can use various tools to facilitate efficient data standardization. The HL7 API (HAPI) FHIR offers a Java API for developing HL7 FHIR clients and servers. Forge serves as a management tool for FHIR profiles. The Firely Terminal, a cross-platform command line tool, provides a wide array of commands for working with FHIR resources and installing and publishing FHIR packages. IG Auto-Builder is another helpful tool that simplifies the creation and publication of IGs, available on the internet [96].

Ultimately, the data standardization stage would generate a set of IGs to ensure consistency and conformity in implementing higher-level services.

Data Management

Various situations can arise in the data management development stage, each bringing different challenges. These situations can fall into 3 options.

The first is to develop an FHIR-native warehouse that the health care IT system vendor manages. In this scenario, the vendor assumes responsibility for designing, implementing, and maintaining the warehouse.

The second is to select a well-established third-party warehouse, such as FHIR Works on Amazon Web Services, IBM FHIR Server, Google Cloud Platform Healthcare API, FHIR Server for Azure, Health Cloud, and Oracle Healthcare Data Repository, to store and explore the FHIR APIs. This approach allows vendors to leverage the capabilities of mature third-party warehouses for FHIR API functionality.

The third is to provide FHIR data using plug-ins. In this scenario, vendors retain their existing data infrastructure and use plug-ins to facilitate data transformation from its native format to the FHIR format. A tool called Facade is available to facilitate this mapping process.

As discussed in the *Data Standardization* section, the discrepancy in granularity between different systems can lead to potential information loss. To mitigate this issue, developers can incorporate a mapping log within the transformer component. When encountering a granularity gap during the mapping process, the mapping log captures and records the lost information, associating it with the corresponding target resource ID. This mapping log serves as a reference for any subsequent services or systems requiring detailed information about the mapping process. If the overlying services need to retrieve the lost information, they can make a request based on the resource ID recorded in the mapping log. This measure allows them to access the details lost during the initial mapping, ensuring that the required information is preserved and available for further analysis or processing.

Ultimately, the data management stage generates a series of FHIR APIs. These APIs serve as a foundation for data exploration and form the backbone of the infrastructure required for high-level services.

Data Integration

Two types of interoperable services are commonly used in the data integration development stage.

The first type is the ISPf, which enables interoperability among multiple organizations. The ISPf comprises 4 key components: gateway, validator, flow control, and log system. The gateway, built by the Spring Cloud Gateway, is responsible for API authorization and forwarding API requests between organizations. The validator ensures that the structure and content of the API data comply with the FHIR object defined in IGs. The HAPI FHIR Validator can build this functionality. The flow control component is designed to limit the number of simultaneous API calls to ensure a stable operation. Redis can effectively fulfill the flow control requirements. As ISPf manages multiple organizations and facilitates data exchange, maintaining a comprehensive log system is crucial for history tracking and auditing. Elasticsearch, a powerful search and analytics engine, can be used to develop the log system within the ISPf, enabling efficient storage and retrieval of API call records.

The second type of interoperable service is represented by apps built by the SMART on FHIR architecture [114]. This architecture consists of 3 key components: the resource server, authorization server, and the SMART on FHIR apps. The resource server is an access layer between the data management layer and the SMART on FHIR apps. The authorization server (an OpenID Connect–compliant web server) authenticates users and issues access tokens. SMART on FHIR apps is designed with specific functionalities and can be substituted based on user preferences.

Use Case

We arbitrarily selected a use case that was in addition to the reviewed articles. Portugal et al [115] designed a smart bed infrastructure with an HIS using FHIR. We mapped it back to the FHIR PG to demonstrate PG's effectiveness and generalizability.

In this case, the roles and responsibilities can be mapped to the FHIR PG–practice design. The authors and their research partners formed an IGs editing group to define IGs consisting of profiles and workflows. The profiles were derived from foundational FHIR resources such as Observation, Device, and ServiceRequest. The workflows defined the frequency at which the smart bed would collect vital signs from the smart bed. Subsequently, the authors' team, acting as a health care IT system vendor, developed a gateway that gathers raw data from sensors and converts it into FHIR for transmission. Although they did not discuss the final applications in detail, it can be

inferred that health care application developers can build a better smart bed monitor based on their infrastructure.

The development architecture described in this paper can also be mapped back to the FHIR PG–development architecture. In the data standardization stage, the authors used the HAPI FHIR for HTTP processing, parsing and serialization, and FHIR REST semantics. It provided a bare-bones structure to build the API. In the data management stage, the authors developed a fog server as a gateway between the smart bed and HIS. This fog server is responsible for collecting raw data from the HIS, transforming it into the FHIR format, and facilitating its integration into the FHIR ecosystem. Finally, in the data integration stage, the authors enabled the HIS software to monitor patient procedures and flows, accompanied by the OAuth2 protocol for secure API communication.

Discussion

Principal Findings

FHIR has shown significant advantages in facilitating interoperability among health IT systems compared with established international standards. However, there are challenges in large-scale implementation and promotion, particularly in different countries. First, countries without incentive policies to encourage FHIR research and implementation may exhibit less enthusiasm for adopting FHIR standards. Second, the lack of a suitable infrastructure to support the implementation process can result in high costs associated with FHIR adoption. Third, the foundational resources provided by FHIR may not directly align with the specific service requirements in different regions, necessitating additional customization processes.

The following steps must be taken to address these challenges. First, it is crucial to have government policies that encourage the evolution and adoption of health care data standards. These policies can stimulate the enthusiasm and investment of stakeholders in the health care ecosystem to promote FHIR implementation on a larger scale. Second, strengthening the infrastructure helps reduce the cost and complexity associated with FHIR adoption, which includes developing services such as FHIR data storage, data standard quality control, and managed services for data operations. Third, FHIR profiles and workflows should be defined to address the specific requirements and characteristics of local health systems. By tailoring FHIR IGs to match the needs of different regions, the gap between FHIR foundational resources and specific service requirements can be bridged.

FHIR holds significant potential in standardizing health care data and promoting service interoperability among health care institutions. Its adoption can drive the transformation of the health care service model and enhance the overall quality of health care services. With the growing recognition of the benefits of FHIR and its demonstrated impact on health care interoperability, more stakeholders are expected to actively participate in enriching its implementation. This collective effort would lead to the emergence of extensive health care service innovations, further enhancing the delivery of high-quality health care services.

Limitations

There are a few current limitations when applying the FHIR PG: (1) PG is derived from the waterfall model that follows a sequential and linear approach. Each step must be completed before proceeding to the next step. Therefore, it is time-consuming and costly to return and modify the previous steps if changes are necessary during the development process. (2) Although PG emphasizes the achievement of interoperability, it leaves out the security discussion. Developers must incorporate additional security mechanisms into PG–development architecture to ensure secure interoperation among multiple organizations.

Conclusions

Owing to the unique characteristics of FHIR, including comprehensive coverage of data definitions, substantial flexibility of data exchange, explicit semantics, and many available open-source tools, FHIR-based services have attracted strong interest from stakeholders in the health care ecosystem. Current studies reveal that many institutions, such as hospitals, regulators, and researchers, have already begun collaborations in actively building FHIR foundational frameworks or application use cases. After conducting the latest literature review, we proposed a general FHIR PG to bridge the gap between FHIR IGs and the practice of building usable services. This PG helps stakeholders identify their participant roles, manage the scope of responsibilities, and develop relevant modules, which we believe would effectively facilitate the application and promotion of HL7 FHIR standards across the health care ecosystem.

Conflicts of Interest

None declared.

References

1. Gold M, McLaughlin C. Assessing HITECH implementation and lessons: 5 years later. Milbank Q 2016 Sep;94(3):654-687 [FREE Full text] [doi: 10.1111/1468-0009.12214] [Medline: 27620687]
2. Protti D. Integrated care, information management and information technology in Canada: have we made any progress in the past 12 years? Healthc Q 2013;16(1):54-59. [Medline: 24863308]
3. Murata C, Yamada T, Chen CC, Ojima T, Hirai H, Kondo K. Barriers to health care among the elderly in Japan. Int J Environ Res Public Health 2010 Apr;7(4):1330-1341 [FREE Full text] [doi: 10.3390/ijerph7041330] [Medline: 20617033]

4. Guiding opinions of the general office of the state council on promoting the construction and development of medical consortiums. China Government Network. 2017. URL: http://www.gov.cn/zhengce/content/2017-04/26/content_5189071.htm [accessed 2022-04-01]
5. Notice on promoting the construction of a compact county-level medical and health community. National Health Commission of the People's Republic of China. 2019 May 28. URL: http://www.nhc.gov.cn/jws/s3580/201905/833cd709c8d346d79dcd774fe81f9d83.shtml [accessed 2022-04-01]
6. World population ageing 2020 highlights. United Nations Department of Economic and Social Affairs. 2020 Oct. URL: https://www.un.org/development/desa/pd/sites/www.un.org.development.desa.pd/files/files/documents/2020/Sep/un_pop_2020_pf_ageing_10_key_messages.pdf [accessed 2022-04-01]
7. FHIR®: it is time to shine for the interoperability standard. Enovacom. URL: https://www.enovacom.com/resource/fhir-it-is-time-to-shine-for-the-interoperability-standard [accessed 2022-04-01]
8. Ayaz M, Pasha MF, Alzahrani MY, Budiarto R, Stiawan D. The Fast Health Interoperability Resources (FHIR) standard: systematic literature review of implementations, applications, challenges and opportunities. JMIR Med Inform 2021 Jul 30;9(7):e21929 [FREE Full text] [doi: 10.2196/21929] [Medline: 34328424]
9. Lehne M, Luijten S, Vom Felde Genannt Imbusch P, Thun S. The use of FHIR in digital health - a review of the scientific literature. Stud Health Technol Inform 2019 Sep 03;267:52-58. [doi: 10.3233/SHTI190805] [Medline: 31483254]
10. Murugan M, Babb LJ, Overby Taylor C, Rasmussen LV, Freimuth RR, Venner E, et al. Genomic considerations for FHIR®; eMERGE implementation lessons. J Biomed Inform 2021 Jun;118:103795 [FREE Full text] [doi: 10.1016/j.jbi.2021.103795] [Medline: 33930535]
11. Seong D, Jung S, Bae S, Chung J, Son DS, Yi B. Fast Healthcare Interoperability Resources (FHIR)-based quality information exchange for clinical next-generation sequencing genomic testing: implementation study. J Med Internet Res 2021 Apr 28;23(4):e26261 [FREE Full text] [doi: 10.2196/26261] [Medline: 33908889]
12. Alterovitz G, Heale B, Jones J, Kreda D, Lin F, Liu L, et al. FHIR Genomics: enabling standardization for precision medicine use cases. NPJ Genom Med 2020;5:13 [FREE Full text] [doi: 10.1038/s41525-020-0115-6] [Medline: 32194985]
13. Klopfenstein SA, Vorisek CN, Shutsko A, Lehne M, Sass J, Löbe M, et al. Fast Healthcare Interoperability Resources (FHIR) in a FAIR metadata registry for COVID-19 research. Stud Health Technol Inform 2021 Nov 18;287:73-77. [doi: 10.3233/SHTI210817] [Medline: 34795084]
14. Khalifa A, Mason CC, Garvin JH, Williams MS, Del Fiol G, Jackson BR, et al. Interoperable genetic lab test reports: mapping key data elements to HL7 FHIR specifications and professional reporting guidelines. J Am Med Inform Assoc 2021 Nov 25;28(12):2617-2625 [FREE Full text] [doi: 10.1093/jamia/ocab201] [Medline: 34569596]
15. Kohli M, Morrison JJ, Wawira J, Morgan MB, Hostetter J, Genereaux B, et al. Creation and curation of the society of imaging informatics in medicine hackathon dataset. J Digit Imaging 2018 Feb 20;31(1):9-12 [FREE Full text] [doi: 10.1007/s10278-017-0003-5] [Medline: 28730549]
16. Madrigal E, Le LP. Digital media archive for gross pathology images based on open-source tools and Fast Healthcare Interoperability Resources (FHIR). Mod Pathol 2021 Sep;34(9):1686-1695 [FREE Full text] [doi: 10.1038/s41379-021-00824-8] [Medline: 34035438]
17. Boufahja A, Nichols S, Pangon V. Custom FHIR resources definition of detailed radiation information for dose management systems. In: Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies - Volume 5: BIOSTEC. 2021 Presented at: BIOSTEC 2021; February 11-13, 2021; Virtual. [doi: 10.5220/0010251104670474]
18. Zong N, Stone DJ, Sharma DK, Wen A, Wang C, Yu Y, et al. Modeling cancer clinical trials using HL7 FHIR to support downstream applications: a case study with colorectal cancer data. Int J Med Inform 2021 Jan;145:104308 [FREE Full text] [doi: 10.1016/j.ijmedinf.2020.104308] [Medline: 33160272]
19. González-Castro L, Cal-González VM, Del Fiol G, López-Nores M. CASIDE: a data model for interoperable cancer survivorship information based on FHIR. J Biomed Inform 2021 Dec;124:103953 [FREE Full text] [doi: 10.1016/j.jbi.2021.103953] [Medline: 34781009]
20. Zong N, Wen A, Stone DJ, Sharma DK, Wang C, Yu Y, et al. Developing an FHIR-based computational pipeline for automatic population of case report forms for colorectal cancer clinical trials using electronic health records. JCO Clin Cancer Inform 2020 Mar 05;4:201-209 [FREE Full text] [doi: 10.1200/CCI.19.00116] [Medline: 32134686]
21. Ludmann D, Pantazoglou E, Otten H. Standardized communication using FHIR and SNOMED CT in treatment of diabetic foot syndrome within the project iFoot. Stud Health Technol Inform 2020 Jun 16;270:1395-1396. [doi: 10.3233/SHTI200459] [Medline: 32570676]
22. Glachs D, Namli T, Jung O, Strohmeier F, Ploessnig M, Rodriguez G. FHIR driven self-management support system for diabetes. Stud Health Technol Inform 2020 Jun 16;270:1291-1292. [doi: 10.3233/SHTI200407] [Medline: 32570624]
23. Bauer DC, Metke-Jimenez A, Maurer-Stroh S, Tiruvayipati S, Wilson LO, Jain Y, et al. Interoperable medical data: the missing link for understanding COVID-19. Transbound Emerg Dis 2021 Jul 29;68(4):1753-1760 [FREE Full text] [doi: 10.1111/tbed.13892] [Medline: 33095970]
24. Sass J, Bartschke A, Lehne M, Essenwanger A, Rinaldi E, Rudolph S, et al. The German Corona Consensus Dataset (GECCO): a standardized dataset for COVID-19 research in university medicine and beyond. BMC Med Inform Decis Mak 2020 Dec 21;20(1):341 [FREE Full text] [doi: 10.1186/s12911-020-01374-w] [Medline: 33349259]

25. Shivers J, Amlung J, Ratanaprayul N, Rhodes B, Biondich P. Enhancing narrative clinical guidance with computer-readable artifacts: authoring FHIR implementation guides based on WHO recommendations. J Biomed Inform 2021 Oct;122:103891 [FREE Full text] [doi: 10.1016/j.jbi.2021.103891] [Medline: 34450285]
26. Benhamida A, Kanas A, Vincze M, Papp KT, Abbassi M, Kozlovszky M. SaECG: a new FHIR Data format revision to enable continuous ECG storage and monitoring. In: Proceedings of the IEEE 20th International Symposium on Computational Intelligence and Informatics (CINTI). 2020 Presented at: IEEE 20th International Symposium on Computational Intelligence and Informatics (CINTI); November 05-07, 2020; Budapest, Hungary. [doi: 10.1109/cinti51262.2020.9305828]
27. Bathelt F, Kümmel M, Helfer S, Kamann C, Sedlmayr M. Formal modelling of FHIR based, medical data exchange using algebraic petri nets. Stud Health Technol Inform 2020 Jun 16;270:597-601. [doi: 10.3233/SHTI200230] [Medline: 32570453]
28. Lenivtceva ID, Kopanitsa G. The pipeline for standardizing Russian unstructured allergy anamnesis using FHIR allergyintolerance resource. Methods Inf Med 2021 Sep 23;60(3-04):95-103. [doi: 10.1055/s-0041-1733945] [Medline: 34425626]
29. Boussadi A, Zapletal E. A Fast Healthcare Interoperability Resources (FHIR) layer implemented over i2b2. BMC Med Inform Decis Mak 2017 Aug 14;17(1):120 [FREE Full text] [doi: 10.1186/s12911-017-0513-6] [Medline: 28806953]
30. Wagholikar KB, Mandel JC, Klann JG, Wattanasin N, Mendis M, Chute CG, et al. SMART-on-FHIR implemented over i2b2. J Am Med Inform Assoc 2017 Mar 01;24(2):398-402 [FREE Full text] [doi: 10.1093/jamia/ocw079] [Medline: 27274012]
31. Jiang G, Kiefer RC, Sharma DK, Prud'hommeaux E, Solbrig HR. A consensus-based approach for harmonizing the OHDSI common data model with HL7 FHIR. Stud Health Technol Inform 2017;245:887-891 [FREE Full text] [Medline: 29295227]
32. Fischer P, Stöhr MR, Gall H, Michel-Backofen A, Majeed RW. Data integration into OMOP CDM for heterogeneous clinical data collections via HL7 FHIR bundles and XSLT. Stud Health Technol Inform 2020 Jun 16;270:138-142. [doi: 10.3233/SHTI200138] [Medline: 32570362]
33. Ladas N, Franz S, Haarbrandt B, Sommer KK, Kohler S, Ballout S, et al. openEHR-to-FHIR: converting openEHR compositions to Fast Healthcare Interoperability Resources (FHIR) for the German Corona Consensus Dataset (GECCO). Stud Health Technol Inform 2022 Jan 14;289:485-486. [doi: 10.3233/SHTI210963] [Medline: 35062196]
34. Fette G, Ertl M, Störk S. Translating openEHR models to FHIR. Stud Health Technol Inform 2020 Jun 16;270:1415-1416. [doi: 10.3233/SHTI200469] [Medline: 32570686]
35. Xiao D, Song C, Nakamura N, Nakayama M. Development of an application concerning fast healthcare interoperability resources based on standardized structured medical information exchange version 2 data. Comput Methods Programs Biomed 2021 Sep;208:106232 [FREE Full text] [doi: 10.1016/j.cmpb.2021.106232] [Medline: 34174764]
36. Dolin RH, Gothi SR, Boxwala A, Heale BS, Husami A, Jones J, et al. vcf2fhir: a utility to convert VCF files into HL7 FHIR format for genomics-EHR integration. BMC Bioinformatics 2021 Mar 02;22(1):104 [FREE Full text] [doi: 10.1186/s12859-021-04039-1] [Medline: 33653260]
37. Peterson KJ, Jiang G, Liu H. A corpus-driven standardization framework for encoding clinical problems with HL7 FHIR. J Biomed Inform 2020 Oct;110:103541 [FREE Full text] [doi: 10.1016/j.jbi.2020.103541] [Medline: 32814201]
38. Wang J, Mathews WC, Pham HA, Xu H, Zhang Y. Opioid2FHIR: a system for extracting FHIR-compatible opioid prescriptions from clinical text. In: Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 2020 Presented at: IEEE International Conference on Bioinformatics and Biomedicine (BIBM); December 16-19, 2020; Seoul, Korea (South). [doi: 10.1109/bibm49941.2020.9313258]
39. Pfaff ER, Champion J, Bradford RL, Clark M, Xu H, Fecho K, et al. Fast Healthcare Interoperability Resources (FHIR) as a meta model to integrate common data models: development of a tool and quantitative validation study. JMIR Med Inform 2019 Oct 16;7(4):e15199 [FREE Full text] [doi: 10.2196/15199] [Medline: 31621639]
40. Lenert LA, Ilatovskiy AV, Agnew J, Rudisill P, Jacobs J, Weatherston D, et al. Automated production of research data marts from a canonical fast healthcare interoperability resource data repository: applications to COVID-19 research. J Am Med Inform Assoc 2021 Jul 30;28(8):1605-1611 [FREE Full text] [doi: 10.1093/jamia/ocab108] [Medline: 33993254]
41. Prud'hommeaux E, Collins J, Booth D, Peterson KJ, Solbrig HR, Jiang G. Development of a FHIR RDF data transformation and validation framework and its evaluation. J Biomed Inform 2021 May;117:103755 [FREE Full text] [doi: 10.1016/j.jbi.2021.103755] [Medline: 33781919]
42. Kiourtis A, Mavrogiorgou A, Kyriazis D. A semantic similarity evaluation for healthcare ontologies matching to HL7 FHIR resources. Stud Health Technol Inform 2020 Jun 16;270:13-17. [doi: 10.3233/SHTI200113] [Medline: 32570337]
43. Demurjian S, Agresta T, Sanzi E, DeStefano J. Alternative approaches for supporting Lattice-based Access Control (LBAC) in the Fast Healthcare Interoperability Resources (FHIR) standard. In: Proceedings of the 16th International Conference on Web Information Systems and Technologies (WEBIST 2020). 2020 Presented at: WEBIST 2020; November 3-5, 2020; Budapest, Hungary URL: http://scitepress.org/Papers/2020/101508/101508.pdf [doi: 10.5220/0010150800930104]
44. Chatterjee A, Pahari N, Prinz A. HL7 FHIR with SNOMED-CT to achieve semantic and structural interoperability in personal health data: a proof-of-concept study. Sensors (Basel) 2022 May 15;22(10):3756 [FREE Full text] [doi: 10.3390/s22103756] [Medline: 35632165]
45. Saripalle R, Sookhak M, Haghparast M. An interoperable UMLS terminology service using FHIR. Future Internet 2020 Nov 16;12(11):199. [doi: 10.3390/fi12110199]

46. Ruminski J, Bujnowski A, Kocejko T, Andrushevich A, Biallas M, Kistler R. The data exchange between smart glasses and healthcare information systems using the HL7 FHIR standard. In: Proceedings of the 9th International Conference on Human System Interactions (HSI). 2016 Presented at: 9th International Conference on Human System Interactions (HSI); July 06-08, 2016; Portsmouth, UK. [doi: 10.1109/hsi.2016.7529684]
47. Saripalle RK. Leveraging FHIR to integrate activity data with electronic health record. Health Technol 2019 Apr 27;10:341-352. [doi: 10.1007/s12553-019-00316-5]
48. Yu J, Kwon SH, Park S, Jun JA, Pyo CS. Design and implementation of real-time bio signals management system based on HL7 FHIR for healthcare services. In: Proceedings of the International Conference on Platform Technology and Service (PlatCon). 2021 Presented at: International Conference on Platform Technology and Service (PlatCon); August 23-25, 2021; Jeju, South Korea. [doi: 10.1109/platcon53246.2021.9680756]
49. Khvastova M, Witt M, Essenwanger A, Sass J, Thun S, Krefting D. Towards interoperability in clinical research - enabling FHIR on the open-source research platform XNAT. J Med Syst 2020 Jul 09;44(8):137 [FREE Full text] [doi: 10.1007/s10916-020-01600-y] [Medline: 32642856]
50. Dridi A, Sassi S, Chbeir R, Faiz S. A flexible semantic integration framework for fully-integrated EHR based on FHIR standard. In: Proceedings of the 12th International Conference on Agents and Artificial Intelligence (ICAART 2020). 2020 Presented at: ICAART 2020; February 22-24, 2020; Valletta, Malta. [doi: 10.5220/0008981506840691]
51. Lee YL, Lee HA, Hsu CY, Kung HH, Chiu HW. Implement an international interoperable PHR by FHIR—a Taiwan innovative application. Sustainability 2020 Dec 28;13(1):198. [doi: 10.3390/su13010198]
52. Tanaka K, Yamamoto R. Implementation of a secured cross-institutional data collection infrastructure by applying HL7 FHIR on an existing distributed EMR storages. Stud Health Technol Inform 2020 Jun 26;272:155-158. [doi: 10.3233/SHTI200517] [Medline: 32604624]
53. Cheng AC, Duda SN, Taylor R, Delacqua F, Lewis AA, Bosler T, et al. REDCap on FHIR: clinical data interoperability services. J Biomed Inform 2021 Sep;121:103871 [FREE Full text] [doi: 10.1016/j.jbi.2021.103871] [Medline: 34298155]
54. Semenov I, Osenev R, Gerasimov S, Kopanitsa G, Denisov D, Andreychuk Y. Experience in developing an FHIR medical data management platform to provide clinical decision support. Int J Environ Res Public Health 2019 Dec 20;17(1):73 [FREE Full text] [doi: 10.3390/ijerph17010073] [Medline: 31861851]
55. Gruendner J, Gulden C, Kampf M, Mate S, Prokosch HU, Zierk J. A framework for criteria-based selection and processing of Fast Healthcare Interoperability Resources (FHIR) data for statistical analysis: design and implementation study. JMIR Med Inform 2021 Apr 01;9(4):e25645 [FREE Full text] [doi: 10.2196/25645] [Medline: 33792554]
56. Cloud Healthcare Pledge. The Information Technology Industry Council. URL: https://www.itic.org/public-policy/CloudHealthcarePledge.pdf [accessed 2022-04-01]
57. Shi W, Giuste FO, Zhu Y, Carpenter AM, Iwinski HJ, Hilton C, et al. A FHIR-compliant application for multi-site and multi-modality pediatric scoliosis patient rehabilitation. In: Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 2021 Presented at: IEEE International Conference on Bioinformatics and Biomedicine (BIBM); December 09-12, 2021; Houston, TX. [doi: 10.1109/bibm52615.2021.9669649]
58. Zampognaro P, Paragliola G, Falanga V. A FHIR based architecture of a multiprotocol IoT Home Gateway supporting dynamic plug of new devices within instrumented environments. In: Proceedings of the IEEE Symposium on Computers and Communications (ISCC). 2021 Presented at: IEEE Symposium on Computers and Communications (ISCC); September 05-08, 2021; Athens, Greece. [doi: 10.1109/iscc53001.2021.9631446]
59. Ploner N, Prokosch HU. Integrating a secure and generic mobile app for patient reported outcome acquisition into an EHR infrastructure based on FHIR resources. Stud Health Technol Inform 2020 Jun 16;270:991-995. [doi: 10.3233/SHTI200310] [Medline: 32570530]
60. Kamel PI, Nagy PG. Patient-centered radiology with FHIR: an introduction to the use of FHIR to offer radiology a clinically integrated platform. J Digit Imaging 2018 Jun 3;31(3):327-333 [FREE Full text] [doi: 10.1007/s10278-018-0087-6] [Medline: 29725963]
61. Mandl KD, Gottlieb D, Mandel JC, Ignatov V, Sayeed R, Grieve G, et al. Push button population health: the SMART/HL7 FHIR bulk data access application programming interface. NPJ Digit Med 2020 Nov 19;3(1):151 [FREE Full text] [doi: 10.1038/s41746-020-00358-4] [Medline: 33299056]
62. Nan J, Xu LQ, Wang Q, Bu C, Ma J, Qiao F. Enabling tiered and coordinated services in a health community of primary care facilities and county hospitals based on HL7 FHIR. In: Proceedings of the IEEE International Conference on Digital Health (ICDH). 2021 Presented at: IEEE International Conference on Digital Health (ICDH); September 05-10, 2021; Chicago, IL. [doi: 10.1109/icdh52753.2021.00048]
63. Taechoyotin P, Prasertsom P, Phanhong M, Wongsutthikoson P, Laohasurayodhin R, Pasuthip N, et al. Health link: scalable health information exchange platform in Thailand. In: Proceedings of the 2nd International Conference on Big Data Analytics and Practices (IBDAP). 2021 Presented at: IBDAP; August 26-27, 2021; Bangkok, Thailand. [doi: 10.1109/ibdap52511.2021.9552033]
64. Maxi K, Morocho V. Integrating medical information software using health level seven and FHIR: a case study. In: Narváez FR, Proaño J, Morillo P, Vallejo D, González Montoya D, Díaz GM, editors. Smart Technologies, Systems and Applications. Cham, Switzerland: Springer; 2022.

65. Rosenau L, Majeed RW, Ingenerf J, Kiel A, Kroll B, Köhler T, et al. Generation of a Fast Healthcare Interoperability Resources (FHIR)-based ontology for federated feasibility queries in the context of COVID-19: feasibility study. JMIR Med Inform 2022 Apr 27;10(4):e35789 [FREE Full text] [doi: 10.2196/35789] [Medline: 35380548]
66. Corici AA, Olaf R, Kraufmann B, Bilig A, Caumanns J, Deglmann M, et al. Interoperable and discrete eHealth data exchange between hospital and patient. In: Proceedings of the 23rd Conference on Innovation in Clouds, Internet and Networks and Workshops (ICIN). 2020 Presented at: 23rd Conference on Innovation in Clouds, Internet and Networks and Workshops (ICIN); February 24-27, 2020; Paris, France. [doi: 10.1109/icin48450.2020.9059335]
67. Papaioannou M, Neocleous A, Savva P, Miguel F, Panayides A, Antoniou Z, et al. A prototype of the national EHR system for Cyprus. Annu Int Conf IEEE Eng Med Biol Soc 2021 Nov;2021:2159-2162. [doi: 10.1109/EMBC46164.2021.9630760] [Medline: 34891716]
68. Hidayat IF, Hermanto BR. A preliminary implementation of HL7 FHIR to achieve interoperability in Indonesia's local EHR. In: Proceedings of the 27th International Conference on Telecommunications (ICT). 2020 Presented at: 27th International Conference on Telecommunications (ICT); October 05-07, 2020; Bali, Indonesia. [doi: 10.1109/ict49546.2020.9239534]
69. Sloane EB, Cooper T, Silva R. MDIRA: IEEE, IHE, and FHIR clinical device and information technology interoperability standards, bridging home to hospital to “hospital-in-home”. In: Proceedings of the SoutheastCon 2021. 2021 Presented at: SoutheastCon 2021; March 10-13, 2021; Atlanta, GA. [doi: 10.1109/southeastcon45413.2021.9401934]
70. Mukhiya SK, Lamo Y. An HL7 FHIR and GraphQL approach for interoperability between heterogeneous Electronic Health Record systems. Health Informatics J 2021 Sep 15;27(3):14604582211043920 [FREE Full text] [doi: 10.1177/14604582211043920] [Medline: 34524029]
71. Gruendner J, Deppenwiese N, Folz M, Köhler T, Kroll B, Prokosch HU, et al. The architecture of a feasibility query portal for distributed COVID-19 Fast Healthcare Interoperability Resources (FHIR) patient data repositories: design and implementation study. JMIR Med Inform 2022 May 25;10(5):e36709 [FREE Full text] [doi: 10.2196/36709] [Medline: 35486893]
72. Gruendner J, Wolf N, Tögel L, Haller F, Prokosch HU, Christoph J. Integrating Genomics and Clinical Data for Statistical Analysis by Using GEnome MINIng (GEMINI) and Fast Healthcare Interoperability Resources (FHIR): system design and implementation. J Med Internet Res 2020 Oct 07;22(10):e19879 [FREE Full text] [doi: 10.2196/19879] [Medline: 33026356]
73. Park CH, You SC, Jeon H, Jeong CW, Choi JW, Park RW. Development and validation of the Radiology Common Data Model (R-CDM) for the international standardization of medical imaging data. Yonsei Med J 2022;63(Suppl):S74. [doi: 10.3349/ymj.2022.63.s74]
74. Ziminski T, Demurjian S, Agresta T. Extending the Fast Healthcare Interoperability Resources (FHIR) with meta resources. In: Proceedings of the 16th International Conference on Software Technologies ICSOFT - Volume 1. 2021 Presented at: ICSOFT 2021; July 6-8, 2021; Online. [doi: 10.5220/0010546501670176]
75. De A, Huang M, Feng T, Yue X, Yao L. Analyzing patient secure messages using a Fast Health Care Interoperability Resources (FIHR)-based data model: development and topic modeling study. J Med Internet Res 2021 Jul 30;23(7):e26770 [FREE Full text] [doi: 10.2196/26770] [Medline: 34328444]
76. Suraj V, Del Vecchio Fitz C, Kleiman LB, Bhavnani SK, Jani C, Shah S, et al. SMART COVID navigator, a clinical decision support tool for COVID-19 treatment: design and development study. J Med Internet Res 2022 Feb 18;24(2):e29279 [FREE Full text] [doi: 10.2196/29279] [Medline: 34932493]
77. Michaels M, Syed S, Lober WB. Blueprint for aligned data exchange for research and public health. J Am Med Inform Assoc 2021 Nov 25;28(12):2702-2706 [FREE Full text] [doi: 10.1093/jamia/ocab210] [Medline: 34613371]
78. Curran RL, Kukhareva PV, Taft T, Weir CR, Reese TJ, Nanjo C, et al. Integrated displays to improve chronic disease management in ambulatory care: a SMART on FHIR application informed by mixed-methods user testing. J Am Med Inform Assoc 2020 Aug 01;27(8):1225-1234 [FREE Full text] [doi: 10.1093/jamia/ocaa099] [Medline: 32719880]
79. Thayer JG, Ferro DF, Miller JM, Karavite D, Grundmeier RW, Utidjian L, et al. Human-centered development of an electronic health record-embedded, interactive information visualization in the emergency department using fast healthcare interoperability resources. J Am Med Inform Assoc 2021 Jul 14;28(7):1401-1410 [FREE Full text] [doi: 10.1093/jamia/ocab016] [Medline: 33682004]
80. Karhade AV, Schwab JH, Del Fiol G, Kawamoto K. SMART on FHIR in spine: integrating clinical prediction models into electronic health records for precision medicine at the point of care. Spine J 2021 Oct;21(10):1649-1651 [FREE Full text] [doi: 10.1016/j.spinee.2020.06.014] [Medline: 32599144]
81. Wesley DB, Blumenthal J, Shah S, Littlejohn RA, Pruitt Z, Dixit R, et al. A novel application of SMART on FHIR architecture for interoperable and scalable integration of patient-reported outcome data with electronic health records. J Am Med Inform Assoc 2021 Sep 18;28(10):2220-2225 [FREE Full text] [doi: 10.1093/jamia/ocab110] [Medline: 34279660]
82. Burkhardt HA, Brandt PS, Lee JR, Karras SW, Bugni PF, Cvitkovic I, et al. StayHome: a FHIR-native mobile COVID-19 symptom tracker and public health reporting tool. Online J Public Health Inform 2021 Mar 21;13(1):e2 [FREE Full text] [doi: 10.5210/ojphi.v13i1.11462] [Medline: 33936522]
83. Hoffman RA, Wu H, Venugopalan J, Braun P, Wang MD. Intelligent mortality reporting with FHIR. In: Proceedings of the IEEE EMBS International Conference on Biomedical & Health Informatics (BHI). 2016 Presented at: IEEE EMBS

International Conference on Biomedical & Health Informatics (BHI); February 16-19, 2017; Orlando, FL. [doi: 10.1109/bhi.2017.7897235]
84. Stoldt JP, Weber JH. Safety improvement for SMART on FHIR apps with data quality by contract. In: Proceedings of the IEEE International Conference on Software Architecture Companion (ICSA-C). 2020 Presented at: IEEE International Conference on Software Architecture Companion (ICSA-C); March 16-20, 2020; Salvador, Brazil. [doi: 10.1109/icsa-c50368.2020.00041]
85. Stoldt JP, Weber JH. Provenance-based trust model for assessing data quality during clinical decision making. In: Proceedings of the IEEE/ACM 3rd International Workshop on Software Engineering for Healthcare (SEH). 2021 Presented at: IEEE/ACM 3rd International Workshop on Software Engineering for Healthcare (SEH); June 3, 2021; Madrid, Spain. [doi: 10.1109/seh52539.2021.00012]
86. Alamri B, Javed IT, Margaria T. A GDPR-compliant framework for IoT-based personal health records using blockchain. In: Proceedings of the 11th IFIP International Conference on New Technologies, Mobility and Security (NTMS). 2021 Presented at: 11th IFIP International Conference on New Technologies, Mobility and Security (NTMS); April 19-21, 2021; Paris, France. [doi: 10.1109/ntms49979.2021.9432661]
87. George M, Chacko AM. A patient-centric interoperable, quorum-based healthcare system for sharing clinical data. In: Proceedings of the 2022 International Conference for Advancement in Technology (ICONAT). 2022 Presented at: 2022 International Conference for Advancement in Technology (ICONAT); January 21-22, 2022; Goa, India. [doi: 10.1109/iconat53423.2022.9725924]
88. Gulden C, Blasini R, Nassirian A, Stein A, Altun FB, Kirchner M, et al. Prototypical clinical trial registry based on Fast Healthcare Interoperability Resources (FHIR): design and implementation study. JMIR Med Inform 2021 Jan 12;9(1):e20470 [FREE Full text] [doi: 10.2196/20470] [Medline: 33433393]
89. Chaves A, Guimarães T, Duarte J, Peixoto H, Abelha A, Machado J. Development of FHIR based web applications for appointment management in healthcare. Proc Comput Sci 2021;184:917-922. [doi: 10.1016/j.procs.2021.03.114]
90. Bae S, Yi BK. Development of eClaim system for private indemnity health insurance in South Korea: compatibility and interoperability. Health Informatics J 2022 Jan 17;28(1):14604582211071019 [FREE Full text] [doi: 10.1177/14604582211071019] [Medline: 35034475]
91. Bettoni GN, Lobo TC, Flores CD, Gomes Tavares dos Santos B, da Silva FP. Application of HL7 FHIR in a microservice architecture for patient navigation on registration and appointments. In: Proceedings of the IEEE/ACM 3rd International Workshop on Software Engineering for Healthcare (SEH). 2021 Presented at: IEEE/ACM 3rd International Workshop on Software Engineering for Healthcare (SEH); June 03, 2021; Madrid, Spain. [doi: 10.1109/seh52539.2021.00015]
92. Weber M, Griessbach A, Grossmann R, Blaser J. A FHIR-based eConsent app for the digital hospital. Stud Health Technol Inform 2020 Jun 16;270:3-7. [doi: 10.3233/SHTI200111] [Medline: 32570335]
93. Sfat R, Marin I, Goga N, Popa R, Darla IC, Marian CV. Conceptualization of an intelligent HL7 application based on questionnaire generation and editing. In: Proceedings of the IEEE International Black Sea Conference on Communications and Networking (BlackSeaCom). 2021 Presented at: IEEE International Black Sea Conference on Communications and Networking (BlackSeaCom); May 24-28, 2021; Bucharest, Romania. [doi: 10.1109/blackseacom52164.2021.9527847]
94. Mohammed S, Fiaidhi J, Sawyer D. Generating physician standing orders for unplanned care scenarios using the HL7 FHIR patient summaries. In: Proceedings of the International Conference on e-Health and Bioengineering (EHB). 2021 Presented at: International Conference on e-Health and Bioengineering (EHB); November 18-19, 2021; Iasi, Romania. [doi: 10.1109/ehb52898.2021.9657715]
95. HAPI FHIR - Java API for HL7 FHIR. GitHub. URL: https://github.com/hapifhir/hapi-fhir [accessed 2023-08-01]
96. Welcome to FHIR®. Health Level Seven International and Fast Healthcare Interoperability Resources. URL: https://build.fhir.org/ [accessed 2023-08-01]
97. Implementation Guide builder. GitHub. URL: https://github.com/FHIR/auto-ig-builder [accessed 2023-08-01]
98. Forge. SIMPLIFIER.NET. URL: https://simplifier.net/forge [accessed 2023-08-01]
99. Firely terminal. SIMPLIFIER.NET. URL: https://simplifier.net/firely-terminal [accessed 2023-08-01]
100. FirelyTeam. GitHub. URL: https://github.com/FirelyTeam/Vonk.Facade.Starter [accessed 2023-08-01]
101. FHIR Works on AWS deployment. GitHub. URL: https://github.com/awslabs/fhir-works-on-aws-deployment [accessed 2023-08-01]
102. FHIR server for Azure. GitHub. URL: https://github.com/microsoft/fhir-server [accessed 2023-08-01]
103. Cloud healthcare API. Google Cloud. URL: https://cloud.google.com/healthcare-api [accessed 2023-08-01]
104. IBM Watson Health is now Merative. IBM. URL: https://ibm.com/products/fhir-server [accessed 2023-08-01]
105. Oracle Healthcare Data Repository homepage. Oracle. URL: https://www.oracle.com/healthcare/data-repository/ [accessed 2023-08-01]
106. Health cloud. Salesforce. URL: https://www.salesforce.com/products/health-cloud/overview/ [accessed 2023-08-01]
107. Spring cloud gateway. GitHub. URL: https://github.com/spring-cloud/spring-cloud-gateway [accessed 2023-08-01]
108. Redis. GitHub. URL: https://github.com/redis/redis [accessed 2023-08-01]
109. validator-wrapper. GitHub. URL: https:/github.com/hapifhir/org.hl7.fhir.validator-wrapper [accessed 2023-08-01]
110. Elasticsearch. GitHub. URL: https://github.com/elastic/elasticsearch [accessed 2023-08-01]

111. OpenID homepage. OpenID. URL: https://openid.net/ [accessed 2023-08-01]
112. OAuth 2.0 homepage. OAuth 2.0. URL: https://oauth.net/ [accessed 2023-08-01]
113. SMART on FHIR. GitHub. URL: https://github.com/smart-on-fhir [accessed 2023-08-01]
114. Mandel JC, Kreda DA, Mandl KD, Kohane IS, Ramoni RB. SMART on FHIR: a standards-based, interoperable apps platform for electronic health records. J Am Med Inform Assoc 2016 Sep;23(5):899-908 [FREE Full text] [doi: 10.1093/jamia/ocv189] [Medline: 26911829]
115. Portugal D, Faria JN, Domingues M, Gaspar L. Integration of a smart bed infrastructure with Hospital Information Systems using Fast Health Interoperability Resources: *a case study of the Wireless biOmonitoring stickers and smart bed architecture: toWards Untethered Patients (WoW) R and D Project. In: Proceedings of the IEEE 20th Consumer Communications & Networking Conference (CCNC). 2023 Presented at: IEEE 20th Consumer Communications & Networking Conference (CCNC); January 08-11, 2023; Las Vegas, NV. [doi: 10.1109/ccnc51644.2023.10060813]

Abbreviations

API: application programming interface
CDA: Clinical Document Architecture
CDM: clinical data model
EHR: electronic health record
EMR: electronic medical record
FHIR: Fast Healthcare Interoperability Resources
HAPI: Health Level 7 application programming interface
HIS: hospital information system
HL7: Health Level 7
IG: Implementation Guide
ISPf: integrated service platform
OMOP: Observational Medical Outcomes Partnership
PG: practice guideline
SMART: Substitutable Medical Applications Reusable Technologies

Edited by M Focsa; submitted 05.12.22; peer-reviewed by C Gulden, S Hume, H Kim, S Sarbadhikari, S Ahalt; comments to author 11.01.23; revised version received 07.04.23; accepted 10.07.23; published 21.08.23.

Please cite as:
Nan J, Xu LQ
Designing Interoperable Health Care Services Based on Fast Healthcare Interoperability Resources: Literature Review
JMIR Med Inform 2023;11:e44842
URL: https://medinform.jmir.org/2023/1/e44842
doi:10.2196/44842
PMID:37603388

Review

Applications of Natural Language Processing for the Management of Stroke Disorders: Scoping Review

Helios De Rosario[1], PhD; Salvador Pitarch-Corresa[1], PT, AD; Ignacio Pedrosa[2], PhD; Marina Vidal-Pedrós[1], PT, MSc; Beatriz de Otto-López[2], PhD; Helena García-Mieres[2], PhD; Lydia Álvarez-Rodríguez[2], MSc

[1]Instituto de Biomecánica de Valencia, Universitat Politècnica de València, Valencia, Spain

[2]CTIC Centro Tecnológico de la Información y la Comunicación, Gijón, Spain

Corresponding Author:
Helios De Rosario, PhD
Instituto de Biomecánica de Valencia
Universitat Politècnica de València
Camino de Vera s/n, Ed. 9C
Valencia, 46022
Spain
Phone: 34 961111170
Email: helios.derosario@ibv.org

Abstract

Background: Recent advances in natural language processing (NLP) have heightened the interest of the medical community in its application to health care in general, in particular to stroke, a medical emergency of great impact. In this rapidly evolving context, it is necessary to learn and understand the experience already accumulated by the medical and scientific community.

Objective: The aim of this scoping review was to explore the studies conducted in the last 10 years using NLP to assist the management of stroke emergencies so as to gain insight on the state of the art, its main contexts of application, and the software tools that are used.

Methods: Data were extracted from Scopus and Medline through PubMed, using the keywords “natural language processing” and “stroke.” Primary research questions were related to the phases, contexts, and types of textual data used in the studies. Secondary research questions were related to the numerical and statistical methods and the software used to process the data. The extracted data were structured in tables and their relative frequencies were calculated. The relationships between categories were analyzed through multiple correspondence analysis.

Results: Twenty-nine papers were included in the review, with the majority being cohort studies of ischemic stroke published in the last 2 years. The majority of papers focused on the use of NLP to assist in the diagnostic phase, followed by the outcome prognosis, using text data from diagnostic reports and in many cases annotations on medical images. The most frequent approach was based on general machine learning techniques applied to the results of relatively simple NLP methods with the support of ontologies and standard vocabularies. Although smaller in number, there has been an increasing body of studies using deep learning techniques on numerical and vectorized representations of the texts obtained with more sophisticated NLP tools.

Conclusions: Studies focused on NLP applied to stroke show specific trends that can be compared to the more general application of artificial intelligence to stroke. The purpose of using NLP is often to improve processes in a clinical context rather than to assist in the rehabilitation process. The state of the art in NLP is represented by deep learning architectures, among which Bidirectional Encoder Representations from Transformers has been found to be especially widely used in the medical field in general, and for stroke in particular, with an increasing focus on the processing of annotations on medical images.

(JMIR Med Inform 2023;11:e48693) doi:10.2196/48693

KEYWORDS

stroke; natural language processing; artificial intelligence; scoping review; scoping; review methods; review methodology; NLP; cardiovascular; machine learning; deep learning

Introduction

Stroke, also called "brain attack," is a medical emergency that occurs when blood flow to a part of the brain is disrupted caused by a clot blocking an artery or by a cerebral hemorrhage due to a ruptured artery. Stroke can result in a range of symptoms and complications depending on the area of the brain that is affected, having impacts on perception, motor control (typically weakness or paralysis on one side of the body, dizziness or difficulty with balance), or behavior (difficulty in speaking or understanding speech), which is a life-threatening emergency that requires immediate medical attention. Although mortality from stroke is decreasing in developed, high-income countries, it remains one of the leading causes of mortality and disability along with ischemic heart disease, and the prevalence of people living with the effects of stroke is increasing due to the growing and aging population [1].

Therefore, the economic and social costs related to the hospitalization, treatment, and recovery of stroke patients are increasing, and there is a growing demand for advanced technologies that can assist in clinical diagnosis, treatment, predictions of clinical events, intervention recommendations, rehabilitation programs, and related factors [2]. For instance, a quick diagnosis and treatment of stroke is crucial as it leads to improved outcomes and prognosis among patients treated within the so-called "golden hour" [3].

In this context, novel approaches that complement and go beyond evidence-based medicine are required. Tools based on artificial intelligence (AI), with their ability to process large amounts of data, have been widely discussed in recent years as one of the proposed approaches to improve the care of stroke, assisting in diagnosis, prognosis, treatment, and prevention [3,4].

AI is an interdisciplinary science with multiple approaches, which in recent years has experienced a significant growth in the fields of machine learning (ML) and deep learning (DL). ML and DL algorithms can learn from data and improve their performance over time without being explicitly programmed, and these methods can deal with very large and complex data sets. DL is considered a recent specialization of ML, which uses artificial neural networks to extract complex representations and features from data. Throughout the manuscript, a distinction is made between DL, used for algorithms based on multilayered neural networks, and traditional ML based on other techniques.

The application of AI to the management of stroke is a topic that has gained a lot of traction in the general field of health informatics [5], partly owing to the remarkable impact of stroke in public health and the subsequent high demand for effective and efficient tools to diagnose and treat stroke. Moreover, the complexity and variety of stroke casuistry make it a good target for AI solutions, which are especially suited to process large amounts of data from a wide range of sources, identify patterns and trends in large data sets, and learn and adapt to new data.

A domain where those advances have produced particularly good results is natural language processing (NLP), which is a promising tool for medicine to unlock the full potential of electronic health records (EHRs), since it might be used to automatically transform clinical text into structured clinical data that can guide clinical decisions [6,7]. The potential of NLP in the analysis of EHR data is particularly appealing given the great quantity of data contained in these records. Notwithstanding their importance, such data are intractable with conventional mathematical methods, since they are recorded in clinical reports, prescriptions, annotations on medical images, and generally unstructured texts [8].

NLP can assist in the identification of patterns and trends in large data sets, which can improve the understanding of factors that contribute to the development of diseases and can in turn help to define more effective prevention and treatment strategies. NLP can also be used in the analysis of particular cases to guide decisions and potentially delay or prevent the onset of the disease. NLP can also be used to develop intelligent systems to find relevant information in the medical literature [9].

Nevertheless, NLP poses particular challenges, including the protection of privacy in the extraction of data, since personal information is often mixed with other data; the variety of the quality and format of EHR data, which depend on the source and software used to collect them; and the difficulty of annotating data samples for training [10]. Therefore, to unlock the potential of NLP in the exploitation of EHRs, researchers and developers need to combine different advanced ML techniques, apply careful data management, and gain a deep understanding of the clinical domain. There is, however, a paucity of guidance on selecting appropriate methods tailored to the health care industry [11].

This scoping review aimed to gather the knowledge that might help in that guidance by investigating how NLP is used to deliver a smarter health care in different phases of stroke disorders (prevention, diagnosis, treatment, and prognosis). The primary questions that served as a guide for the review are: (1) In which phases or contexts of stroke management is NLP used (prevention, diagnosis, treatment, and/or prognosis)? (2) Which are the main benefits of applying NLP to stroke management, related to clinical, social, and economic factors? and (3) What types of clinical data are collected and used by NLP in stroke management (ie, demographic data, medical notes, physical and functional examination, reports of laboratory or medical devices)?

This review also focused on the following secondary questions: (1) What NLP methods, AI algorithms, and tools are used in stroke studies? (2) Which AI techniques or frameworks are used to process and analyze the data? (3) Are there algorithms and NLP software specifically tuned for stroke? and (4) Which tools have the best performance and how do they compare to others?

Methods

Design

The unregistered protocol for this review was created following the PRISMA-ScR (Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews) guidelines [12] and the JBI Manual for Scoping Reviews [13].

Inclusion Criteria

The target patient population of this scoping review included adults that had suffered stroke and people at risk of stroke due to a history of predisposing vascular background or other conditions that increase the risk of developing stroke, including mental illness or heart diseases such as a reduced ejection fraction.

The main concept of interest was the use of NLP in stroke management in public or private health care systems, including use cases and the data and technologies involved in those applications. We considered both the application of NLP for monitoring and decision-making of individual patients as well as for the planification of care resources in the management of stroke cases.

We were interested in any context where prevention, treatment, or rehabilitation of stroke might take place, ranging from early detection outside or inside clinical settings, diagnosis and evaluation of cases, clinical decision-making, administration and monitoring of rehabilitation, and postrehabilitation management.

The types of evidence sources taken into account included articles from peer-reviewed journals, books, and conference papers, considering both primary research studies and systematic or scoping reviews, as well as reports from scientific, medical, or government institutions.

Search Strategy

The search was performed in the electronic databases of Scopus and Medline through PubMed, using the keywords “natural language processing” and “stroke,” restricted to articles published in the last 10 years, between 2013 and 2022.

Selection Process

The results of the search were imported into the Zotero Reference Manager software (Corporation for Digital Scholarship, Virginia), which was used to filter out duplicate records. Titles and abstracts of the filtered list were screened independently by two reviewers to ascertain their eligibility according to the inclusion criteria. Disagreements were resolved in a discussion session between the reviewers to obtain a consensus.

The full text of the papers was read by two independent reviewers to extract the relevant data as described below. An internal cross-validation by three other experts on the topic was also considered. Works whose content did not meet the eligibility criteria or did not contain sufficient information to answer the primary questions were excluded and those that reported the same results from the same study were treated as duplicates. The record of rejected works was shared by the reviewers to confirm the decisions of either part.

Data Extraction and Presentation of Results

The reviewers filled out a table with the following data from each work included in the final selection: type of study, primary diagnosis, related diseases that were used either as inclusion criteria or as predictors in the data analysis, sample size (if suitable), and qualitative responses to the primary and secondary questions.

Works were classified depending on whether or not they reported experimental studies, and those that did were further subclassified as clinical trials or different types of observational studies: cross-sectional, retrospective or prospective, and cohort or case-control studies.

A dictionary of terms was defined for the tabulated records of the primary and secondary questions and their relative frequencies were calculated. In addition, the relationships between answers were analyzed in two different multiple correspondence analyses (MCAs), which can be employed to detect and represent underlying structures in categorical data sets (ie, frequent co-occurrence of specific categories in two or more variables) [14]. One of the MCAs focused on the primary questions, seeking relationships between the context of application (eg, classification of diagnostics, prognosis of outcomes) and the types of data that were processed. The other MCA focused on the secondary questions, seeking relationships between NLP methods and software tools. In both analyses, the type of AI models (general ML, DL, or rule-based algorithms) was also included as a variable. The analysis was performed in R [15], using the packages *factoMineR* [16] and *factoextra* [17] for MCA and its graphical representation.

Results

General Description of the Studies

A total of 115 unique papers were identified out of 223 records obtained in the search; 29 studies were eventually included for data extraction and analysis after screening by title and abstract and reading of the full text (see the flow diagram in Figure 1).

The general characteristics of the 29 reviewed studies (year, type of study, target diseases, and sample size), together with the items extracted from the primary and secondary questions are respectively presented in Tables 1, 2, and 3.

Figure 1. Flow diagram of the review process. NLP: natural language processing.

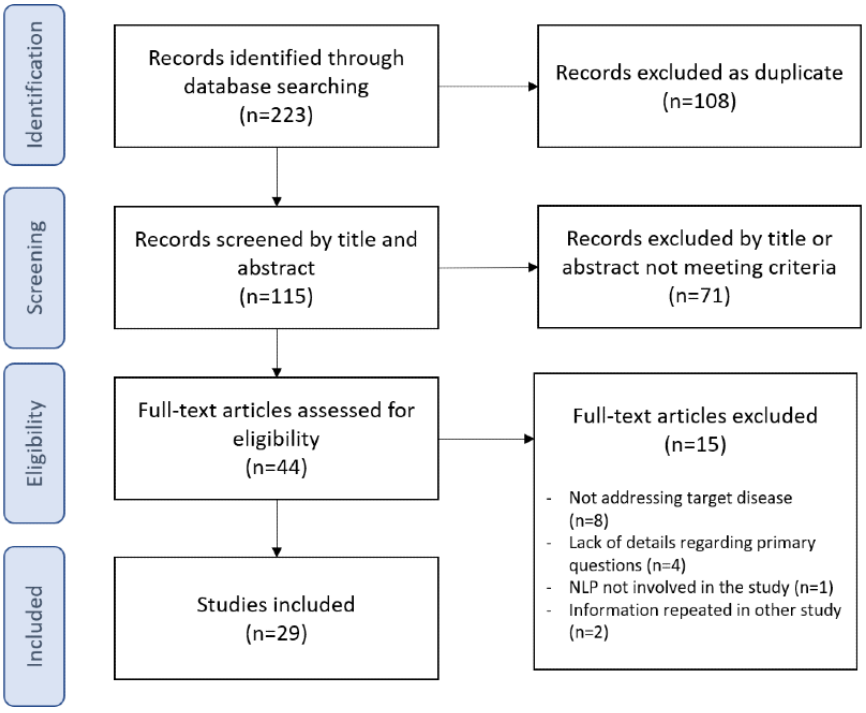


Table 1. Summary of the included studies: study type, sample size, type of stroke, and other diseases or conditions taken into account.

Reference	Year	Type of study	Sample size[a]	Type of stroke	Other conditions
Zhao et al [18]	2021	Cohort study	4914	Transient ischemic attack, hemorrhagic stroke	AF[b]
Zanotto et al [19]	2021	Retrospective cross-sectional cohort study	188	Ischemic stroke	AF, CAD[c], DM[d], dyslipidemia, hypertension, smoking, other[e]
Sung et al [20]	2022	Retrospective cohort study	3847	Acute ischemic stroke	AF, CHF[f], DM, cancer, hyperlipidemia, hypertension
Sung et al [21]	2021	Retrospective cohort study	3847	Acute ischemic stroke	AF, CHF, DM, cancer, hyperlipidemia, hypertension
Miller et al [22]	2022	Retrospective cohort study	918	Ischemic stroke	Other
Mayampurath et al [23]	2021	Cohort study	965	Acute ischemic stroke, hemorrhagic stroke	Other
Lineback et al [24]	2021	Retrospective cohort study	2855	Ischemic stroke, hemorrhagic stroke	AF, CAD, CHF, DM, cancer, hyperlipidemia, hypertension, other
Kogan et al [25]	2020	Retrospective cohort study	7149	Ischemic stroke, hemorrhagic stroke, transient ischemic attack	None
Heo et al [26]	2020	Retrospective cohort study	1810	Acute ischemic stroke	DM, dyslipidemia, hyperglycemia, hypertension, smoking, other
Deng et al [27]	2022	Feasibility study	1000 (simulated)	Hemorrhagic stroke	DM, hypertension
Bacchi et al [28]	2019	Cohort study	2201	Transient ischemic attack	None
Yu et al [29]	2021	Cohort study	1320	Ischemic stroke, hemorrhagic stroke	None
Wheater et al [30]	2019	Cohort study	2160	Ischemic stroke, hemorrhagic stroke	None
Sung et al [31]	2020	Cohort study	4640	Acute ischemic stroke	None
Sung et al [32]	2018	Feasibility study	90	Acute ischemic stroke	Hyperglycemia, other
Shek et al [33]	2021	Cohort study	2327	Stroke comorbidities	AF, CHF, DM, hypertension
Rannikmäe et al [34]	2021	Cohort study	207	Intracerebral hemorrhage, subarachnoid hemorrhage, and ischemic stroke	None
Ong et al [35]	2020	Cohort study	721	Acute ischemic stroke	None
Mowery et al [36]	2016	Cohort study	498	Ischemic stroke	CAD, CHF, DM, hypertension
Li et al [37]	2021	Cohort study	3971	Acute or subacute ischemic stroke	None
Leung et al [38]	2021	Cohort study	182	Not applicable	Other
Kim et al [39]	2019	Cohort study	3204	Acute ischemic stroke	None
Kent et al [40]	2021	Retrospective cohort study	261,960	Ischemic stroke	AF, CAD, CHF, DM, hyperlipidemia, hypertension, other
Lin et al [41]	2021	Retrospective cohort study	1700	Acute ischemic stroke	Other
Guan et al [42]	2021	Cohort study	1598	Ischemic stroke	CHF, other
Garg et al [43]	2019	Cohort study	1091	Ischemic stroke	AF, CAD, DM, hyperlipidemia, hypertension
Farran et al [44]	2022	Retrospective cohort study	16,916	Not applicable	AF
Elkin et al [45]	2021	Cohort study	96,681	Not applicable	AF
Bacchi et al [46]	2022	Cohort study	438	Ischemic stroke, hemorrhagic stroke	None

[a]Number of patients involved.

[b]AF: atrial fibrillation.

[c]CAD: coronary artery disease.

[d]DM: diabetes mellitus.

[e]Other refers to conditions that are not already listed in the table.

[f]CHF: coronary heart failure.

The vast majority were cohort studies that analyzed clinical aspects, along with societal or economic aspects of the disease in some cases, at the moment of data gathering. Approximately one third of the papers (n=10) also included a retrospective analysis and 2 of them were limited to feasibility studies. Although the search included a time span of 10 years, only one of the studies included in the review was older than 5 years [36] and most studies (n=19) had been published in the last 2 years (2021 or 2022).

Most studies (n=24) focused on ischemic stroke (either acute, subacute, or transient); the second most frequent type of stroke was hemorrhagic stroke (n=9), which in the majority of cases was in addition to and not excluding ischemic stroke (only 2 papers dealt exclusively with hemorrhagic stroke). Many studies considered other clinical conditions that were used to select the patients or were included as information taken into account by the models. The most common conditions were atrial fibrillation, diabetes mellitus, and hypertension; each of them was considered in one third of the reviewed papers (n=10). Other diseases that were considered with smaller frequency were hyper- or dyslipidemia, hyperglycemia, hypercholesterolemia, coronary heart failure, smoking, or cancer.

The sample size of the cohort studies was highly varied, ranging between 182 patients [38] and more than 260,000 patients [40], with a median sample size of 2160 patients. The two feasibility studies were conducted either with simulated cases [27] or with a smaller sample of 90 patients [32].

Table 4 shows the frequency of each category used to classify the answers to the primary and secondary questions, except for the question about the specificity of algorithms and NLP tools for stroke, since there was little variability in those answers.

Table 2. Summary of the answers to the primary questions.

Reference	Context for NLP[a] use	Expected benefits	Types of clinical data[b]
Zhao et al [18]	Prevention and diagnosis (classification)	CLINICAL: improved triage	Demographic data, laboratory test results, medical history, medication
Zanotto et al [19]	Prognosis (outcomes)	CLINICAL: care information management, characterize patients, prediction of outcomes, risk assessment; SOCIETAL: supporting research studies; ECONOMIC: public health management	Diagnostic reports
Sung et al [20]	Prognosis (outcomes)	CLINICAL: prediction of outcomes	Annotated medical images, clinical scales, demographic data, diagnostic reports, medical history, patient treatments
Sung et al [21]	Prognosis (outcomes)	CLINICAL: prediction of outcomes, risk assessment	Annotated medical images, clinical scales, demographic data, diagnostic reports, functional outcomes data
Miller et al [22]	Prognosis (outcomes)	CLINICAL: prediction of outcomes, risk assessment	Annotated medical images, diagnostic reports
Mayampurath et al [23]	Diagnosis (classification)	CLINICAL: improved triage	Diagnostic reports
Lineback et al [24]	Prognosis (recurrence)	CLINICAL: care information management	Demographic data, diagnostic reports, medical history, medication, patient treatments
Kogan et al [25]	Prognosis (outcomes)	CLINICAL: administration of treatments, care information management, improved triage, prediction of outcomes	Demographic data, clinical scales, medical history, patient treatments, medication
Heo et al [26]	Prognosis (outcomes)	CLINICAL: prediction of outcomes	Annotated medical images, diagnostic reports
Deng et al [27]	Diagnosis (details); treatment	CLINICAL: administration of treatments	Annotated medical images, clinical scales, diagnostic reports, medical history
Bacchi et al [28]	Diagnosis (classification)	CLINICAL: stroke cause prediction	Annotated medical images, diagnostic reports, medical history, medication
Yu et al [29]	Diagnosis (details)	CLINICAL: improved triage; ECONOMIC: public health management	Annotated medical images, diagnostic reports
Wheater et al [30]	Diagnosis (classification)	CLINICAL: disease surveillance, improved triage; ECONOMIC: public health management	Annotated medical images, diagnostic reports
Sung et al [31]	Prevention and diagnosis (classification)	CLINICAL: administration of treatments, care information management, disease surveillance; ECONOMIC: public health management	Diagnostic reports
Sung et al [32]	Diagnosis (details); treatment	CLINICAL: administration of treatments	Diagnostic reports, laboratory test results, medical history
Shek et al [33]	Diagnosis (comorbidities)	CLINICAL: care information management	Demographic data, medical history
Rannikmäe et al [34]	Diagnosis (classification)	CLINICAL: improved triage	Annotated medical images, diagnostic reports
Ong et al [35]	Diagnosis (details)	CLINICAL: administration of treatments, prediction of outcomes; SOCIETAL: supporting research studies	Annotated medical images, diagnostic reports
Mowery et al [36]	Prevention	CLINICAL: risk assessment	Diagnostic reports
Li et al [37]	Diagnosis (classification)	CLINICAL: improved triage	Annotated medical images, diagnostic reports
Leung et al [38]	Diagnosis (details)	CLINICAL: care information management, characterize patients	Annotated medical images, diagnostic reports
Kim et al [39]	Diagnosis (classification)	CLINICAL: care information management, characterize patients	Annotated medical images, laboratory results, demographic data, diagnostic reports, functional outcomes data

Reference	Context for NLP[a] use	Expected benefits	Types of clinical data[b]
Kent et al [40]	Prognosis (outcomes)	CLINICAL: care information management, characterize patients, stroke cause prediction	Annotated medical images, diagnostic reports
Lin et al [41]	Diagnosis (details); prognosis (recurrence)	SOCIETAL: supporting research studies	Diagnostic reports
Guan et al [42]	Diagnosis (classification)	CLINICAL: improved triage	Clinical scales, diagnostic reports
Garg et al [43]	Diagnosis (classification)	CLINICAL: improved triage, risk assessment	Annotated medical images, diagnostic reports, medical history
Farran et al [44]	Diagnosis (classification); prognosis (outcomes)	CLINICAL: stroke cause prediction, disease surveillance; ECONOMIC: public health management	Clinical scales, demographic data, medical history, patient treatments
Elkin et al [45]	Diagnosis (classification)	Not applicable	Clinical scales, demographic data
Bacchi et al [46]	Diagnosis (classification)	Not applicable	diagnostic reports, patient treatment

[a]NLP: natural language processing.

[b]See Multimedia Appendix 1 for the definitions of clinical data types, following Jiang et al [6].

Table 3. Summary of the answers to the secondary questions.

Reference	AI[a] technique	NLP[b] methods[c]	Other statistical methods[c]	Software packages[c,d]	Performance metrics[c]	Best performing methods
Zhao et al [18]	ML[e]	Regular expressions	LR[f], RF[g]	MedTagger, Weka	PPV[h], NPV[i], F1, sensitivity	RF
Zanotto et al [19]	ML	Ontologies (OWL[j]), BERT[k], BOW[l], TF-IDF[m]	CNN[n], K-NN[o], RF, SVM[p], naïve Bayes	spaCy	PPV, F1, sensitivity	SVM ontological rules
Sung et al [20]	ML	Negation extraction ontologies (UMLS[q])	Gradient boosting	Jazzy spell checker, MetaMap, XGBoost[r]	AUC[s], IDI[t], NRI[u]	Not applicable
Sung et al [21]	DL	BOW, BERT (ClinicalBERT)	Not applicable	Jazzy spell checker	AUC, IDI, NRI	Not applicable
Miller et al [22]	DL rule-based	BOW, negation extraction, TF-IDF, BERT (BioClinicalBERT)	LASSO[v], K-NN, RF, MLP[w]	scikit-learn	AUC, PPV, sensitivity, specificity	BioClinicalBERT (except for rare and continuous outcomes)
Mayampurath et al [23]	ML	N-grams (1- or 2-)	SVM	Not applicable	AUC, PPV, NPV, sensitivity, specificity	Not applicable
Lineback et al [24]	ML	N-grams (1- or 2-), TF-IDF, Word-embedding (Word2Vec)	LASSO, LR, PCA[x], RF, SVM, gradient boosting, naïve Bayes	XGBoost	AUC	ML methods in general
Kogan et al [25]	ML rule-based	Not applicable	RF, gradient boosting, MLP	Not applicable	Correlations, RMSE[y]	Not applicable
Heo et al [26]	DL	BOW, Word-embedding (sent2vec, BioWordVec)	Decision trees, CNN, LASSO, LSTM[z], MLP, RF, SVM	Quanteda, NLTK[aa], Tensorflow, Keras	AUC	Document-level methods, CNN
Deng et al [27]	DL rule-based	BERT	Not applicable	Not applicable	AUC, PPV, NPV, sensitivity, specificity	Not applicable
Bacchi et al [28]	DL	BOW, negation extraction	Decision trees, CNN, LSTM, RF	Not applicable	AUC, PPV, NPV, sensitivity, specificity	CNN
Yu et al [29]	Rule-based	Regular expressions	Not applicable	CHARTextract	PPV, NPV, accuracy, sensitivity, specificity	Not applicable
Wheater et al [30]	Rule-based	Regular expressions, grammatical analysis, ontologies (custom), negation extraction	Not applicable	BRAT rapid annotation tool	PPV, sensitivity, specificity	Not applicable
Sung et al [31]	ML rule-based	Grammatical analysis (part-of-speech), negation extraction, ontologies (UMLS)	Decision trees (CART[bb]), K-NN, LR, RF, SVM	Google spell checker, MetaMap, Weka	Accuracy, κ	Mixed results
Sung et al [32]	Not applicable	Grammatical analysis (part-of-speech), negation extraction, ontologies (UMLS)	Not applicable	Google spell checker, MetaMap, Stata	NPV, F1, sensitivity, specificity	Document-level methods
Shek et al [33]	DL	Grammatical analysis, Negation extraction, Ontologies (SNOMED[cc])	Not applicable	MedCAT	NPV, F1, sensitivity, specificity	Not applicable
Rannikmäe et al [34]	ML rule-based	Ontologies (UMLS)	Not applicable	SemEHR	PPV, sensitivity	Mixed results
Ong et al [35]	DL	BOW, TF-IDF, Word-embedding (GloVE[dd])	Decision trees (CART), K-NN, LR, LSTM, RF	scikit-learn, Tensorflow	AUC, F1, accuracy, sensitivity, specificity	GloVE + LSTM
Mowery et al [36]	Rule-based	Regular expressions	Not applicable	pyConTeXT	PPV, NPV, sensitivity, specificity	Not applicable
Li et al [37]	ML	BOW, N-gram (2- and 3-), negation extraction	RF	scikit-learn, NLTK	F1, accuracy	Not applicable

Reference	AI[a] technique	NLP[b] methods[c]	Other statistical methods[c]	Software packages[c,d]	Performance metrics[c]	Best performing methods
Leung et al [38]	DL rule-based	Not applicable	Not applicable	MedTagger	PPV, NPV, accuracy, sensitivity, specificity	Not applicable
Kim et al [39]	ML	N-gram (1- and 2-), TF-IDF	Decision trees, LR, naïve Bayes, RF, SVM	Quanteda	AUC, F1	Single decision trees
Kent et al [40]	DL rule-based	Ontologies (named entity recognition)	Not applicable	MedTagger	PPV, NPV, accuracy, sensitivity, specificity	Not applicable
Lin et al [41]	DL	BERT (ClinicalBERT, StrokeBERT)	Not applicable	spaCy	AUC, F1	StrokeBERT
Guan et al [42]	ML	Regular expressions, negation extraction	Decision trees (CART), K-NN, LR, RF, SVM	Quanteda	AUC, PPV, NPV, F1, accuracy, specificity	RF
Garg et al [43]	ML	BOW, N-grams (1- to 3-)	Decision trees, K-NN, stacking LR, PCA, RF, SVM, gradient boosting	cTAKES, spaCy, XGBoost	AUC, sensitivity, κ	Stacking, LR, gradient boost
Farran et al [44]	ML	Ontologies (SNOMED), negation extraction	Not applicable	MedCAT	Accuracy	Not applicable
Elkin et al [45]	ML	Ontologies (SNOMED)	Not applicable	HD-NLP[ee]	PPV, NPV, sensitivity, specificity	Not applicable
Bacchi et al [46]	ML	BOW, N-grams (1- to 3-), negation extraction	Decision trees, LR, RF	scikit-learn, NLTK	AUC, PPN, NPP, sensitivity, specificity	RF

[a]AI: artificial intelligence.

[b]NLP: natural language processing.

[c]See brief descriptions of the NLP tools, statistical methods, software packages, and performance metrics in Multimedia Appendix 2 [47-51].

[d]Excluding general programming frameworks like Python or R.

[e]ML: machine learning.

[f]LR: logistic regression.

[g]RF: random forest.

[h]PPV: positive predictive value.

[i]NPV: negative predictive value.

[j]OWL: Web Ontology Language.

[k]BERT: Bidirectional Encoder Representations from Transformers.

[l]BOW: bag-of-words.

[m]TF-IDF: term frequency-inverse document frequency.

[n]CNN: convolutional neural network.

[o]K-NN: K-nearest neighbor.

[p]SVM: support vector machine.

[q]UMLS: Unified Medical Language System.

[r]XGBoost: extreme gradient boosting.

[s]AUC: area under the curve.

[t]IDI: integrated discrimination index.

[u]NRI: Net Reclassification Index.

[v]LASSO: least absolute shrinkage and selection operator.

[w]MLP: multilayer perceptron.

[x]PCA: principal component analysis.

[y]RMSE: root mean squared error.

[z]LSTM: long short-term memory.

[aa]NLTK: Natural Language Processing toolkit for Python.

[bb]CART: classification and regression tree.

[cc]SNOMED: Systematized Nomenclature of Medicine.

[dd]GLoVE: Global Vectors for Word Representation.

[ee]HD-NLP: high-definition natural language processing.

Table 4. Frequencies of distinctive items found in primary and secondary questions among the included studies (N=29).[a]

Variable and category[b]	Studies, n (%)
Context	
Diagnostic (classification)	13 (45)
Diagnostic (details)	6 (21)
Prognostic (outcomes)	8 (28)
Prognostic (recurrence)	2 (7)
Prevention	3 (10)
Treatment	2 (7)
Clinical benefits	
Improved triage	9 (31)
Care information management	8 (28)
Prediction of outcomes	7 (24)
Administration of treatments	5 (17)
Risk assessment	5 (17)
Patient characterization	4 (14)
Disease surveillance	3 (10)
Stroke causes	3 (10)
Data sources	
Diagnostic reports	24 (83)
Annotated images	15 (52)
Medical history	10 (34)
Demographic data	9 (31)
Clinical scales	7 (24)
Treatments	5 (17)
Medication	4 (14)
Laboratory results	3 (10)
Functional outcomes data	2 (7)
Artificial intelligence technique	
ML[c]	15 (52)
DL[d]	10 (34)
Rule-based	10 (34)
Natural language processing tools	
Negation extraction (NEGEX)	11 (38)
Ontologies	10 (34)
Bag-of-words (BOW)	
n-grams	6 (21)
Bidirectional Encoder Representations from Transformers (BERT)	5 (17)
Regular expressions (REG-EXPR)	5 (17)
TF-IDF[e]	5 (17)
Grammatical analysis	4 (14)
Word-embedding	3 (10)
Other statistical tools	

Variable and category[b]	Studies, n (%)
Random forest (RF)	14 (48)
Decision trees	8 (28)
Support vector machine (SVM)	7 (24)
Logistic regression (LR)	7 (24)
K-nearest neighbor (K-NN)	6 (21)
Gradient boosting	4 (14)
Naïve Bayes	3 (10)
Multilayer perceptron (MLP)	3 (10)
Long short-term memory (LSTM)	3 (10)
Principal component analysis (PCA)	2 (7)
Software packages	
scikit-learn	4 (14)
NLTK[f]	3 (10)
spaCy	3 (10)
Quanteda	3 (10)
MedTagger	3 (10)
MetaMap	3 (10)
XGBoost[g]	3 (10)
MedCAT	2 (7)
Weka	2 (7)
Tensorflow	2 (7)
Performance metrics	
Based on ratios (PPV[h], NPV[i], F1, accuracy, sensitivity, or specificity)	23 (79)
Based on ROC[j] curves (AUC[k], C-statistic)	14 (48)
Differential measures (NRI[l], IDI[m])	2 (7)

[a]Only the items that occurred more than once are reported in this table; however, since different items often overlapped in each study, the frequencies of each variable normally sum to more than 100%.

[b]See brief descriptions of the NLP tools, statistical methods, software packages, and performance metrics in Multimedia Appendix 2 [47-51].

[c]ML: machine learning.

[d]DL: deep learning.

[e]TF-IDF: term frequency-inverse document frequency.

[f]NLTK: Natural Language Processing toolkit for Python.

[g]XGBoost: extreme gradient boosting.

[h]PPV: positive predictive value.

[i]NPV: negative predictive value.

[j]ROC: receiver operating characteristic.

[k]AUC: area under the curve.

[l]NRI: Net Reclassification Index.

[m]IDI: integrated discrimination index.

The most frequent context of stroke in which the studies were applied was the diagnostic phase, followed by the prognosis of outcomes. The potential benefit of the results on clinical processes (eg, improving the triage of patients depending on the type or severity of stroke, more efficient management of care information) was the main focus of all studies but one [41], which chiefly focused on the societal aspect of supporting research studies, similar to two other studies that also evaluated that aspect along with clinical applications. Five of the 29 studies (17%) also considered the potential economic benefit of NLP, in terms of reducing the costs of stroke for the public health sector.

The most frequent source of data for NLP models was diagnostic reports (n=24), followed in many cases by annotations on medical images such as radiographs and scans (n=15). General ML models were used more frequently than DL or rule-based algorithms to process the data (n=15 for ML vs n=10 papers for either DL or rule-based techniques). NLP tools, other statistical methods, and the software packages that were used to implement them highly varied across papers, although there were some associations with the AI technique and other variables (see the next subsection).

In nearly all studies, the AI architectures and algorithms had been adapted to deal with stroke-related data, except for one study that used an ML model for patients with severe mental illness at risk of stroke [44]. One of the studies actually used a software tool that was specifically designed for stroke [41], StrokeBERT, which is a language representation model based on Google's Bidirectional Encoder Representations from Transformers (BERT) [47]. Other studies used models that were adapted to broader medical terminology, including ClinicalBERT [52], BioClinicalBERT [53], and BioWordVec [54], or models tuned with standard medical vocabularies such as Systematized Nomenclature of Medicine (SNOMED) [55] or Unified Medical Language System (UMLS) [56].

The methods used to compare the performance of the models were also highly varied, although in the greatest majority of cases (n=23) they were metrics based on the ratios of true/false-positive or -negative values (positive predictive value, negative predictive value, sensitivity, specificity, F1 score, or accuracy), and many were based on the receiver operating characteristic curve (n=14); a few studies (n=2) also used measures of classification improvements such as the net reclassification index and the integrated discrimination index [48], and only one study used other statistics such as correlation coefficients or the root mean squared error [25].

Owing to the variety of methods and tools used in the studies, there were few coincidences in the selection of the best ones. The only methods that were chosen as the best performing in more than one study were random forest (n=3), convolutional neural network (n=2), and BERT (n=2).

Multiple Correspondence Analysis

Figures 2 and 3 show the proximity of the categories that exhibited the closest relationships in the two first dimensions obtained in the MCA.

The common variable used in the analysis (AI technique) was clearly distinguished in the first two dimensions of the MCA plot, which on the one hand separated rule-based techniques from ML and DL and on the other hand separated general ML from DL.

In the first MCA (Figure 2), it could be observed that the studies focusing on the classification of diagnostics (often used for the triage of patients) and prospects of recurrent stroke were often those that also used ML techniques with demographic data and information on treatments. Although the other categories were less tightly related, the text associated with clinical tests and the annotations on images were related more closely to prognostics of outcomes than to other contexts of application, with annotated images also being used to ascertain details of the stroke episode. Both types of studies were frequently approached by DL and sometimes by rule-based techniques.

In the other MCA (Figure 3), AI techniques were separated between ML, DL, and rule-based methods in the two main dimensions of the projected space, although only general ML and DL were closely related to other items.

Figure 2. Projection of the scores of the categories in the first two dimensions of the multiple correspondence analysis plot involving context of application, data sources, and artificial intelligence technique. DL: deep learning; ML: machine learning.

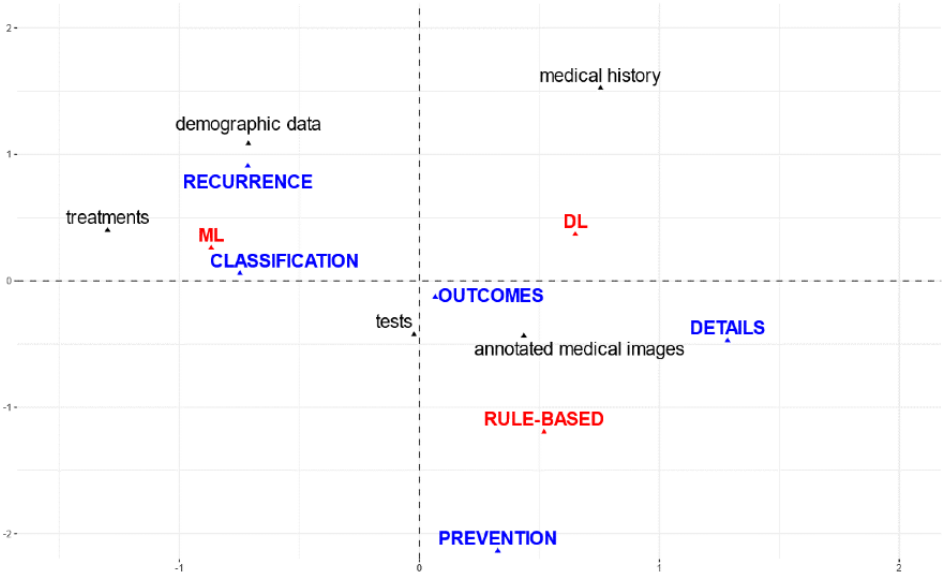
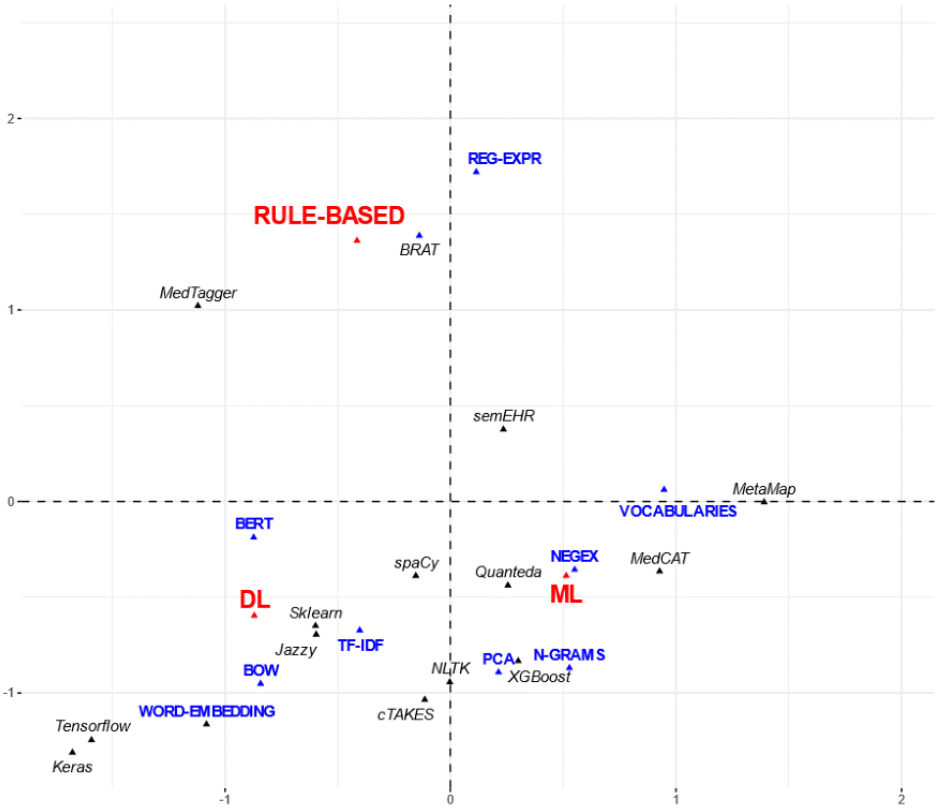


Figure 3. Projection of the scores of the categories in the first two dimensions of the multiple correspondence analysis plot involving natural language processing methods, software, and artificial intelligence techniques. See brief descriptions of the methods and software in Multimedia Appendix 2. BERT: Bidirectional Encoder Representations from Transformers; BOW: Bag-of-words; BRAT: Browser-based Rapid Annotation Tool; DL: deep learning; ML: machine learning; NEGEX: Negation extraction; NLTK: Natural Language Processing toolkit for Python; REG-EXPR; regular expressions; TF-IDF; term frequency-inverse document frequency; XGBoost: extreme gradient boosting.



ML was related to NLP methods that are used in the first steps of the processing pipeline, such as the extraction of text tokens in the form of *n*-grams, detection of negated terms, and use of standard vocabularies. This was mostly performed with software tools such as MetaMap, MedCAT, Quanteda, and extreme gradient boosting.

Conversely, DL was more associated with the usage of BERT, a language representation model based on transformers [47], and NLP methods applied to numerical and vectorized representations of the language tokens, such as the "bag-of-words," term frequency-inverse document frequency word embeddings, and other word embeddings. This was chiefly performed with software packages such as Tensorflow through Keras and scikit-learn. Other software packages that are often used for NLP, such as Natural Language Processing toolkit for Python, were observed in the middle of the primary axis of the MCA plot, halfway between the general ML and DL architectures.

Discussion

The research on AI for stroke management has gained greater interest and impact in the last few years [5], and the growing rate of publications found in this scoping review reveals that the same trend is occurring in research on NLP, which is a particular field of AI, applied to the same clinical condition. However, in other aspects, the studies focused on NLP show their own specific trends.

Although the search for this scoping review was very broad, and did not limit the type and phase of stroke to be studied, the vast majority of studies were focused on ischemic stroke in its acute, subacute, or transient stage, and the purpose of using NLP was to improve processes in a clinical context. This focus on clinical contexts is related to the relevance that is attributed to the unstructured information contained in EHRs, (ie, in notes, reports, and annotated images) as predictors of outcomes and complications, which are crucial for proper decision-making, together with the difficulty of processing that information automatically with traditional tools. The deployment of NLP models integrated in the pipelines of an EHR, programmed to automatically ingest and process incoming records [57], or even the patients' commentaries in emergency through voice-to-text [58], may be used to identify patients at high risk and requiring prompt access to specific treatments; find signs to anticipate impending stroke; or evaluate its severity, type, and risks of complications.

Efficient triage of patients in emergency and early consultations, more accurate diagnostics, or prognostics of outcomes and recurrence were the main intended applications of NLP models in the reviewed studies. Accordingly, the main sources of information exploited by NLP algorithms were clinical data of the patients obtained from their history, especially the diagnostic reports of the current stroke episode. Administration and monitoring of rehabilitation, or postrehabilitation management, were not dealt with in the final selection of studies that were the object of the review.

NLP is itself a broad concept, which involves many types of computational techniques. In its more general sense, NLP comprises all methods and tools that can be used to analyze texts in order to represent human languages, based either on theory of language constructs, semantic mappings, or emulation of linguistic processes occurring in the human brain [59]. The relationships between these tools, types of statistical and ML models, data sources, and applications found by the MCA help to understand how each subset of techniques can be used to solve different problems, and can also help to interpret some trends in the evolution of this technology applied to the clinical management of stroke.

Some of these methods rely on text-processing algorithms that use predefined rules and vocabularies, such as the tokenization of long texts into smaller items, categorization of those items in parts of speech, and construction of syntactic structures, and they have been widely used since long before the recent revolution of big data and DL fields. What this revolution has provided to the field of NLP is the maturity of more complex representations of language data, such as the word embeddings into large-dimensional numeric vectors and their effective processing through deep neural networks, as well as the exploitation of huge databases of texts, such as the Common Crawl data set that includes petabytes of text data, crawled monthly from dozens of billions of web pages [60].

In this context, the state of the art in NLP is represented by DL architectures such as GPT, XLNet, or BERT [61]. Among these, BERT has been found to be particularly widely used in the medical field in general, and for stroke in particular, along with specialized versions fitted to these applications that improve their performance [22,41]. More basic ML algorithms and hybrid approaches with rule-based techniques are still more present than advanced DL networks in the recent research on NLP for stroke, and in some cases, tailored rule-based systems outperformed BERT and its derivatives [19,22]. Support vector machine methods were also found to perform better than BERT in one study [19], although random forest was reported to have the best performance more frequently than any other ML method in the set of reviewed studies [18,42,46]. Some of these results may seem unexpected, given the remarkable performance of DL in general, and particularly large language models (LLMs), in other areas. However, the computational complexity and large data sets needed to train LLMs can limit their current scalability, not outperforming other ML methods that work better on limited training data such as the data sets of the mentioned studies.

The prevalence of studies based on traditional ML methods over those that use DL neural networks may be partly due to the recency of the more complex DL architectures, as well as to the need of larger sets of data to train those models, which raises the bar to conduct studies with that approach. However, it is also interesting to observe that the choice of the AI technique also relates to the type of data that are processed and the context of application of NLP, such that DL is more closely related to studies that involve medical imaging with annotations to prognosticate the outcomes of stroke.

Taking into account these pieces of evidence, and considering the future of NLP in stroke, further development of LLMs in the biomedical field may be expected. LLMs emerged in 2018 as a class of language models that use neural networks with billions of parameters trained on huge amounts of unlabeled text data through self-supervised learning. LLMs are often based on transformers, a self-attention mechanism to compute contextual relationships between the input tokens [62]. However, innovation in the NLP field will come from the development of these models for medical specialties such as stroke. These biomedical LLMs can be trained not only with data sources from EHRs but also from scientific and clinical publications and social network posts from specialized fields. The particularity is that these models need to be trained on much larger databases than those used by classical ML algorithms to achieve adequate performance metrics. This involves combining computational resources and very large data sources, an option that is not always available for the existing resources in research.

Acknowledgments

This review was conducted within the framework of the IBERUS project Technological Network of Biomedical Engineering Applied to Degenerative Pathologies of the Neuromusculoskeletal System in Clinical and Outpatient Settings (CER-20211003) and the CERVERA Network financed by the Ministry of Science and Innovation through the Center for Industrial Technological Development (CDTI), charged to the General State Budgets 2021 and the Recovery, Transformation, and Resilience Plan.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Categories of clinical data.
[DOCX File , 15 KB - medinform_v11i1e48693_app1.docx]

Multimedia Appendix 2
Description of artificial intelligence (AI), natural language processing (NLP), and statistical tools.
[DOCX File , 20 KB - medinform_v11i1e48693_app2.docx]

Multimedia Appendix 3
PRISMA-ScR checklist.
[PDF File (Adobe PDF File), 103 KB - medinform_v11i1e48693_app3.pdf]

References

1. Stinear CM, Lang CE, Zeiler S, Byblow WD. Advances and challenges in stroke rehabilitation. Lancet Neurol 2020 Apr;19(4):348-360. [doi: 10.1016/S1474-4422(19)30415-6] [Medline: 32004440]
2. Sirsat MS, Fermé E, Câmara J. Machine learning for brain stroke: a review. J Stroke Cerebrovasc Dis 2020 Oct;29(10):105162. [doi: 10.1016/j.jstrokecerebrovasdis.2020.105162] [Medline: 32912543]
3. Abedi V, Khan A, Chaudhary D, Misra D, Avula V, Mathrawala D, et al. Using artificial intelligence for improving stroke diagnosis in emergency departments: a practical framework. Ther Adv Neurol Disord 2020 Aug 25;13:1756286420938962 [FREE Full text] [doi: 10.1177/1756286420938962] [Medline: 32922515]
4. Thompson MP, Fanaroff AC, Parker JD, Vallabhajosyula S, Sterling MR. Focusing on the future of cardiovascular outcomes research: highlights From the American Heart Association/American Stroke Association Quality of Care and Outcomes Research 2018 Scientific Sessions. Circ Cardiovasc Qual Outcomes 2018 Jun;11(6):e004871. [doi: 10.1161/CIRCOUTCOMES.118.004871] [Medline: 29903937]
5. Luvizutto GJ, Silva GF, Nascimento MR, Sousa Santos KC, Appelt PA, de Moura Neto E, et al. Use of artificial intelligence as an instrument of evaluation after stroke: a scoping review based on international classification of functioning, disability and health concept. Top Stroke Rehabil 2022 Jul 11;29(5):331-346. [doi: 10.1080/10749357.2021.1926149] [Medline: 34115576]
6. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol 2017 Dec;2(4):230-243 [FREE Full text] [doi: 10.1136/svn-2017-000101] [Medline: 29507784]
7. Sheikhalishahi S, Miotto R, Dudley JT, Lavelli A, Rinaldi F, Osmani V. Natural language processing of clinical notes on chronic diseases: systematic review. JMIR Med Inform 2019 Apr 27;7(2):e12239 [FREE Full text] [doi: 10.2196/12239] [Medline: 31066697]
8. Adnan K, Akbar R, Khor S, Ali ABA. Role and challenges of unstructured big data in healthcare. In: Sharma N, Chakrabarti A, Balas VE, editors. Data management, analytics and innovation. Advances in intelligent systems and computing. Singapore: Springer; 2020:301-323.
9. Sneiderman CA, Rindflesch TC, Aronson AR. Finding the findings: identification of findings in medical literature using restricted natural language processing. Proc AMIA Annu Fall Symp 1996:239-243 [FREE Full text] [Medline: 8947664]
10. Li I, Pan J, Goldwasser J, Verma N, Wong WP, Nuzumlalı MY, et al. Neural natural language processing for unstructured data in electronic health records: a review. Comput Sci Rev 2022 Nov;46:100511. [doi: 10.1016/j.cosrev.2022.100511]
11. Shahid N, Rappon T, Berta W. Applications of artificial neural networks in health care organizational decision-making: a scoping review. PLoS One 2019;14(2):e0212356 [FREE Full text] [doi: 10.1371/journal.pone.0212356] [Medline: 30779785]
12. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
13. Peters M, Godfrey C, McInerney P, Munn Z, Tricco A, Khalil H. Chapter 11: Scoping reviews. In: Aromataris E, Munn Z, editors. JBI Manual for Evidence Synthesis. Adelaide, Australia: JBI Collaboration; 2020.
14. Husson F, Josse J. Multiple correspondence analysis. In: Blasius J, Greenacre M, editors. Visualization and verbalization of data. Boca Raton, FL: Chapman and Hall/CRC; 2014.
15. R Core Team. R: A Language and Environment for Statistical Computing. 2020. URL: http://www.R-project.org/ [accessed 2022-12-12]
16. Lê S, Josse J, Husson F. FactoMineR: an R package for multivariate analysis. J Stat Soft 2008;25(1):1-18. [doi: 10.18637/jss.v025.i01]
17. Kassambara A, Mundt F. Factoextra: extract and visualize the results of multivariate data analyses. CRAN R Project. 2020. URL: https://CRAN.R-project.org/package=factoextra [accessed 2022-12-12]
18. Zhao Y, Fu S, Bielinski SJ, Decker PA, Chamberlain AM, Roger VL, et al. Natural language processing and machine learning for identifying incident stroke from electronic health records: algorithm development and validation. J Med Internet Res 2021 Mar 08;23(3):e22951 [FREE Full text] [doi: 10.2196/22951] [Medline: 33683212]
19. Zanotto BS, Beck da Silva Etges AP, Dal Bosco A, Cortes EG, Ruschel R, De Souza AC, et al. Stroke outcome measurements from electronic medical records: cross-sectional study on the effectiveness of neural and nonneural classifiers. JMIR Med Inform 2021 Nov 01;9(11):e29120 [FREE Full text] [doi: 10.2196/29120] [Medline: 34723829]
20. Sung S, Hsieh C, Hu Y. Early prediction of functional outcomes after acute ischemic stroke using unstructured clinical text: retrospective cohort study. JMIR Med Inform 2022 Feb 17;10(2):e29806 [FREE Full text] [doi: 10.2196/29806] [Medline: 35175201]
21. Sung S, Chen C, Pan R, Hu Y, Jeng J. Natural language processing enhances prediction of functional outcome after acute ischemic stroke. J Am Heart Assoc 2021 Dec 21;10(24):e023486 [FREE Full text] [doi: 10.1161/JAHA.121.023486] [Medline: 34796719]

22. Miller MI, Orfanoudaki A, Cronin M, Saglam H, So Yeon Kim I, Balogun O, et al. Natural language processing of radiology reports to detect complications of ischemic stroke. Neurocrit Care 2022 Aug 09;37(Suppl 2):291-302 [FREE Full text] [doi: 10.1007/s12028-022-01513-3] [Medline: 35534660]
23. Mayampurath A, Parnianpour Z, Richards CT, Meurer WJ, Lee J, Ankenman B, et al. Improving prehospital stroke diagnosis using natural language processing of paramedic reports. Stroke 2021 Aug;52(8):2676-2679 [FREE Full text] [doi: 10.1161/STROKEAHA.120.033580] [Medline: 34162217]
24. Lineback CM, Garg R, Oh E, Naidech AM, Holl JL, Prabhakaran S. Prediction of 30-day readmission after stroke using machine learning and natural language processing. Front Neurol 2021 Jul 13;12:649521 [FREE Full text] [doi: 10.3389/fneur.2021.649521] [Medline: 34326805]
25. Kogan E, Twyman K, Heap J, Milentijevic D, Lin JH, Alberts M. Assessing stroke severity using electronic health record data: a machine learning approach. BMC Med Inform Decis Mak 2020 Jan 08;20(1):8 [FREE Full text] [doi: 10.1186/s12911-019-1010-x] [Medline: 31914991]
26. Heo TS, Kim YS, Choi JM, Jeong YS, Seo SY, Lee JH, et al. Prediction of stroke outcome using natural language processing-based machine learning of radiology report of brain MRI. J Pers Med 2020 Dec 16;10(4):286 [FREE Full text] [doi: 10.3390/jpm10040286] [Medline: 33339385]
27. Deng B, Zhu W, Sun X, Xie Y, Dan W, Zhan Y, et al. Development and validation of an automatic system for intracerebral hemorrhage medical text recognition and treatment plan output. Front Aging Neurosci 2022 Apr 8;14:798132 [FREE Full text] [doi: 10.3389/fnagi.2022.798132] [Medline: 35462698]
28. Bacchi S, Zerner T, Oakden-Rayner L, Kleinig T, Patel S, Jannes J. Deep learning in the prediction of ischaemic stroke thrombolysis functional outcomes: a pilot study. Acad Radiol 2020 Feb;27(2):e19-e23. [doi: 10.1016/j.acra.2019.03.015] [Medline: 31053480]
29. Yu AYX, Liu ZA, Pou-Prom C, Lopes K, Kapral MK, Aviv RI, et al. Automating stroke data extraction from free-text radiology reports using natural language processing: instrument validation study. JMIR Med Inform 2021 May 04;9(5):e24381 [FREE Full text] [doi: 10.2196/24381] [Medline: 33944791]
30. Wheater E, Mair G, Sudlow C, Alex B, Grover C, Whiteley W. A validated natural language processing algorithm for brain imaging phenotypes from radiology reports in UK electronic health records. BMC Med Inform Decis Mak 2019 Sep 09;19(1):184 [FREE Full text] [doi: 10.1186/s12911-019-0908-7] [Medline: 31500613]
31. Sung S, Lin C, Hu Y. EMR-based phenotyping of ischemic stroke using supervised machine learning and text mining techniques. IEEE J Biomed Health Inform 2020 Oct;24(10):2922-2931. [doi: 10.1109/jbhi.2020.2976931]
32. Sung S, Chen K, Wu DP, Hung L, Su Y, Hu Y. Applying natural language processing techniques to develop a task-specific EMR interface for timely stroke thrombolysis: A feasibility study. Int J Med Inform 2018 Apr;112:149-157. [doi: 10.1016/j.ijmedinf.2018.02.005] [Medline: 29500013]
33. Shek A, Jiang Z, Teo J, Au Yeung J, Bhalla A, Richardson MP, et al. Machine learning-enabled multitrust audit of stroke comorbidities using natural language processing. Eur J Neurol 2021 Dec 29;28(12):4090-4097. [doi: 10.1111/ene.15071] [Medline: 34407269]
34. Rannikmäe K, Wu H, Tominey S, Whiteley W, Allen N, Sudlow C, et al. Developing automated methods for disease subtyping in UK Biobank: an exemplar study on stroke. BMC Med Inform Decis Mak 2021 Jun 15;21(1):191 [FREE Full text] [doi: 10.1186/s12911-021-01556-0] [Medline: 34130677]
35. Ong CJ, Orfanoudaki A, Zhang R, Caprasse FPM, Hutch M, Ma L, et al. Machine learning and natural language processing methods to identify ischemic stroke, acuity and location from radiology reports. PLoS One 2020 Jun 19;15(6):e0234908 [FREE Full text] [doi: 10.1371/journal.pone.0234908] [Medline: 32559211]
36. Mowery DL, Chapman BE, Conway M, South BR, Madden E, Keyhani S, et al. Extracting a stroke phenotype risk factor from Veteran Health Administration clinical reports: an information content analysis. J Biomed Semantics 2016 May 10;7(1):26 [FREE Full text] [doi: 10.1186/s13326-016-0065-1] [Medline: 27175226]
37. Li M, Lang M, Deng F, Chang K, Buch K, Rincon S, et al. Analysis of stroke detection during the COVID-19 pandemic using natural language processing of radiology reports. AJNR Am J Neuroradiol 2021 Mar;42(3):429-434 [FREE Full text] [doi: 10.3174/ajnr.A6961] [Medline: 33334851]
38. Leung LY, Fu S, Luetmer PH, Kallmes DF, Madan N, Weinstein G, et al. Agreement between neuroimages and reports for natural language processing-based detection of silent brain infarcts and white matter disease. BMC Neurol 2021 May 11;21(1):189 [FREE Full text] [doi: 10.1186/s12883-021-02221-9] [Medline: 33975556]
39. Kim C, Zhu V, Obeid J, Lenert L. Natural language processing and machine learning algorithm to identify brain MRI reports with acute ischemic stroke. PLoS One 2019;14(2):e0212778 [FREE Full text] [doi: 10.1371/journal.pone.0212778] [Medline: 30818342]
40. Kent DM, Leung LY, Zhou Y, Luetmer PH, Kallmes DF, Nelson J, et al. Association of silent cerebrovascular disease identified using natural language processing and future ischemic stroke. Neurology 2021 Sep 28;97(13):e1313-e1321 [FREE Full text] [doi: 10.1212/WNL.0000000000012602] [Medline: 34376505]
41. Lin C, Hsu K, Liang C, Lee T, Liou C, Lee J, et al. A disease-specific language representation model for cerebrovascular disease research. Comput Methods Programs Biomed 2021 Nov;211:106446 [FREE Full text] [doi: 10.1016/j.cmpb.2021.106446] [Medline: 34627022]

42. Guan W, Ko D, Khurshid S, Trisini Lipsanopoulos AT, Ashburner JM, Harrington LX, et al. Automated electronic phenotyping of cardioembolic stroke. Stroke 2021 Jan;52(1):181-189 [FREE Full text] [doi: 10.1161/STROKEAHA.120.030663] [Medline: 33297865]
43. Garg R, Oh E, Naidech A, Kording K, Prabhakaran S. Automating ischemic stroke subtype classification using machine learning and natural language processing. J Stroke Cerebrovasc Dis 2019 Jul;28(7):2045-2051. [doi: 10.1016/j.jstrokecerebrovasdis.2019.02.004] [Medline: 31103549]
44. Farran D, Bean D, Wang T, Msosa Y, Casetta C, Dobson R, et al. Anticoagulation for atrial fibrillation in people with serious mental illness in the general hospital setting. J Psychiatr Res 2022 Sep;153:167-173 [FREE Full text] [doi: 10.1016/j.jpsychires.2022.06.044] [Medline: 35816976]
45. Elkin PL, Mullin S, Mardekian J, Crowner C, Sakilay S, Sinha S, et al. Using artificial intelligence with natural language processing to combine electronic health record's structured and free text data to identify nonvalvular atrial fibrillation to decrease strokes and death: evaluation and case-control study. J Med Internet Res 2021 Nov 09;23(11):e28946 [FREE Full text] [doi: 10.2196/28946] [Medline: 34751659]
46. Bacchi S, Gluck S, Koblar S, Jannes J, Kleinig T. Automated information extraction from free-text medical documents for stroke key performance indicators: a pilot study. Intern Med J 2022 Feb 20;52(2):315-317. [doi: 10.1111/imj.15678] [Medline: 35187820]
47. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv. 2019 May 24. URL: https://arxiv.org/abs/1810.04805 [accessed 2022-12-12]
48. Pencina MJ, D'Agostino RB, D'Agostino RB, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med 2008 Jan 30;27(2):157-72; discussion 207. [doi: 10.1002/sim.2929] [Medline: 17569110]
49. Harkema H, Dowling JN, Thornblade T, Chapman WW. ConText: an algorithm for determining negation, experiencer, and temporal status from clinical reports. J Biomed Inform 2009 Oct;42(5):839-851 [FREE Full text] [doi: 10.1016/j.jbi.2009.05.002] [Medline: 19435614]
50. Resnick MP, LeHouillier F, Brown SH, Campbell KE, Montella D, Elkin PL. Automated modeling of clinical narrative with high definition natural language processing using Solor and Analysis Normal Form. Stud Health Technol Inform 2021 Nov 18;287:89-93 [FREE Full text] [doi: 10.3233/SHTI210822] [Medline: 34795088]
51. Wu H, Toti G, Morley KI, Ibrahim ZM, Folarin A, Jackson R, et al. SemEHR: A general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research. J Am Med Inform Assoc 2018 May 01;25(5):530-537 [FREE Full text] [doi: 10.1093/jamia/ocx160] [Medline: 29361077]
52. Huang K, Altosaar J, Ranganath R. ClinicalBERT: modeling clinical notes and predicting hospital readmission. arXiv. 2020 Nov 29. URL: https://arxiv.org/abs/1904.05342 [accessed 2022-12-12]
53. Alsentzer E, Murphy J, Boag W, Weng WH, Jindi D, Naumann T, et al. Publicly available clinical BERT embeddings. 2019 Presented at: 2nd Clinical Natural Language Processing Workshop; June 2019; Minneapolis, MN. [doi: 10.18653/v1/w19-1909]
54. Zhang Y, Chen Q, Yang Z, Lin H, Lu Z. BioWordVec, improving biomedical word embeddings with subword information and MeSH. Sci Data 2019 May 10;6(1):52. [doi: 10.1038/s41597-019-0055-0] [Medline: 31076572]
55. Stearns MQ, Price C, Spackman KA, Wang AY. SNOMED clinical terms: overview of the development process and project status. Proc AMIA Symp 2001:662-666 [FREE Full text] [Medline: 11825268]
56. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004 Jan 01;32(Database issue):D267-D270 [FREE Full text] [doi: 10.1093/nar/gkh061] [Medline: 14681409]
57. Afshar M, Sharma B, Dligach D, Oguss M, Brown R, Chhabra N, et al. Development and multimodal validation of a substance misuse algorithm for referral to treatment using artificial intelligence (SMART-AI): a retrospective deep learning study. Lancet Digit Health 2022 Jun;4(6):e426-e435 [FREE Full text] [doi: 10.1016/S2589-7500(22)00041-3] [Medline: 35623797]
58. Cho A, Min IK, Hong S, Chung HS, Lee HS, Kim JH. Effect of applying a real-time medical record input assistance system with voice artificial intelligence on triage task performance in the emergency department: prospective interventional study. JMIR Med Inform 2022 Aug 31;10(8):e39892 [FREE Full text] [doi: 10.2196/39892] [Medline: 36044254]
59. Chowdhary KR. Natural language processing. In: Fundamentals of artificial intelligence. India: Springer; 2020:603-649.
60. Patel JM. Introduction to common crawl datasets. In: Getting structured data from the internet: running web crawlers/scrapers on a big data production scale. New York: Apress; 2020:277-324.
61. Topal MO, Bas A, van Heerden I. Exploring transformers in natural language generation: GPT, BERT, and XLNet. arXiv. 2021 Feb 16. URL: https://arxiv.org/abs/2102.08036 [accessed 2022-12-12]
62. Fan L, Li L, Ma Z, Lee S, Yu H, Hemphill L. A bibliometric review of large language models research from 2017 to 2023. arXiv. 2023 Apr 03. URL: https://arxiv.org/abs/2304.02020 [accessed 2023-08-03]

Abbreviations

AI: artificial intelligence

BERT: Bidirectional Encoder Representations from Transformers
DL: deep learning
EHR: electronic health record
LLM: large language model
MCA: multiple correspondence analysis
ML: machine learning
NLP: natural language processing
PRISMA-ScR: Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews
SNOMED: Systematized Nomenclature of Medicine
UMLS: Unified Medical Language System

Edited by C Lovis; submitted 03.05.23; peer-reviewed by J Heo, SF Sung; comments to author 05.06.23; revised version received 26.07.23; accepted 28.07.23; published 06.09.23.

Please cite as:
De Rosario H, Pitarch-Corresa S, Pedrosa I, Vidal-Pedrós M, de Otto-López B, García-Mieres H, Álvarez-Rodríguez L
Applications of Natural Language Processing for the Management of Stroke Disorders: Scoping Review
JMIR Med Inform 2023;11:e48693
URL: https://medinform.jmir.org/2023/1/e48693
doi:10.2196/48693
PMID:37672328

Review

Artificial Intelligence–Based Methods for Integrating Local and Global Features for Brain Cancer Imaging: Scoping Review

Hazrat Ali[1], PhD; Rizwan Qureshi[2], PhD; Zubair Shah[1], PhD

[1]College of Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar

[2]Department of Imaging Physics, MD Anderson Cancer Center, University of Texas, Houston, Houston, TX, United States

Corresponding Author:
Zubair Shah, PhD
College of Science and Engineering
Hamad Bin Khalifa University
Al Luqta St
Ar-Rayyan
Doha, 34110
Qatar
Phone: 974 50744851
Email: zshah@hbku.edu.qa

Abstract

Background: Transformer-based models are gaining popularity in medical imaging and cancer imaging applications. Many recent studies have demonstrated the use of transformer-based models for brain cancer imaging applications such as diagnosis and tumor segmentation.

Objective: This study aims to review how different vision transformers (ViTs) contributed to advancing brain cancer diagnosis and tumor segmentation using brain image data. This study examines the different architectures developed for enhancing the task of brain tumor segmentation. Furthermore, it explores how the ViT-based models augmented the performance of convolutional neural networks for brain cancer imaging.

Methods: This review performed the study search and study selection following the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines. The search comprised 4 popular scientific databases: PubMed, Scopus, IEEE Xplore, and Google Scholar. The search terms were formulated to cover the interventions (ie, ViTs) and the target application (ie, brain cancer imaging). The title and abstract for study selection were performed by 2 reviewers independently and validated by a third reviewer. Data extraction was performed by 2 reviewers and validated by a third reviewer. Finally, the data were synthesized using a narrative approach.

Results: Of the 736 retrieved studies, 22 (3%) were included in this review. These studies were published in 2021 and 2022. The most commonly addressed task in these studies was tumor segmentation using ViTs. No study reported early detection of brain cancer. Among the different ViT architectures, Shifted Window transformer–based architectures have recently become the most popular choice of the research community. Among the included architectures, UNet transformer and TransUNet had the highest number of parameters and thus needed a cluster of as many as 8 graphics processing units for model training. The brain tumor segmentation challenge data set was the most popular data set used in the included studies. ViT was used in different combinations with convolutional neural networks to capture both the global and local context of the input brain imaging data.

Conclusions: It can be argued that the computational complexity of transformer architectures is a bottleneck in advancing the field and enabling clinical transformations. This review provides the current state of knowledge on the topic, and the findings of this review will be helpful for researchers in the field of medical artificial intelligence and its applications in brain cancer.

(JMIR Med Inform 2023;11:e47445) doi:10.2196/47445

KEYWORDS

artificial intelligence; AI; brain cancer; brain tumor; medical imaging; segmentation; vision transformers

Introduction

Background

Brain cancer is typically characterized by a brain tumor. A brain tumor is a mass or development of aberrant brain cells. The signs and symptoms of a brain tumor vary widely and are determined by the size, location, and rate of growth of the brain tumor. Brain tumors can originate in the brain (primary brain tumors) or move from other body regions to the brain (secondary or metastatic brain tumors). In general, studying brain cancer is challenging given the highly complex anatomy of the human brain, where several sections are responsible for various nervous system processes [1].

Medical imaging technologies for studying the brain are rapidly advancing. Therefore, it is critical to provide tools to extract information from brain image data such that they may aid in automatic or semiautomatic computer-aided diagnosis of brain cancer. Artificial intelligence (AI) techniques based on modern machine learning and deep learning models enable computers to make data-driven predictions using massive amounts of data. These techniques have a wide range of applications, many of which can be customized to extract useful information from medical images [2-6].

Among AI techniques developed for brain cancer applications, architectures based on convolutional neural networks (CNNs) have dominated the research on brain cancer diagnosis and classification. For example, UNet (an encoder-decoder CNN architecture) and its variants [7,8] are popular for brain tumor segmentation tasks. However, CNNs are known to be effective in extracting only local dependencies in the input image data, which is mainly attributed to the localized receptive field. Compared with CNNs, attention-based transformer models (transformers) [9] are good at capturing long-range dependencies. Given their ability to learn long-range dependencies, transformers form the backbone of most state-of-the-art models in the natural language processing domain [10].

For image classification tasks, Dosovitskiy et al [11] proposed the computer vision variants of the transformer architecture, typically known as vision transformer (ViT). The concept of attention was applied to images by representing them as a sequential combination of 16×16-pixel patches. The image patches were processed in a way similar to tokens (words) in natural language processing [11]. The sections (with positional embeddings) are ordered. The embeddings are vectors that can be learned. Each piece is organized in a straight line and multiplied by the embedding matrix. The position embedding result is passed to the transformer encoder.

Given the potential demonstrated by transformer-based approaches for computer vision tasks, transformers have quickly penetrated the field of medical imaging. For example, some studies [12-15] have used them on computed tomography scans and x-ray images of the lungs to classify COVID-19 and pneumonia. Similarly, Zhang and Zhang [16] and Xie et al [17] used ViT for medical image segmentation, and He et al [18] used ViT for brain age estimation. With the recent developments of ViTs in computer vision applications, there has been a growing interest in developing ViT-based architectures for cancer imaging applications. ViT can also aid in the diagnosis and prognosis of other types of cancers. For example, Chen et al [19] showed the scaling of ViTs to large whole-slide imaging for 33 different cancer types. The benchmarking results demonstrate that the transformer-based architecture with hierarchical pretraining outperforms the existing cancer subtyping and survival prediction methods, indicating its effectiveness in capturing the hierarchical phenotypic structure in tumor microenvironments.

Accordingly, many recent efforts have been reported on the developments of ViT architectures to make progress in brain cancer applications. With the growing interest in developing ViT-based methods for brain cancer imaging, there is a dire need to review the recent developments and identify the key challenges. To the best of our knowledge, no study (review) has reported the different ViT architectures for brain cancer imaging and analyzed how ViT complements CNNs in brain cancer diagnosis, classification, grading, and brain tumor segmentation.

A few review and survey articles that are relevant to our work are by Parvaiz et al [20], Magadza and Viriri [21], Akinyelu et al [22], He et al [23], and Biratu et al [24]. Among these, Magadza and Viriri [21] and Biratu et al [24] have surveyed the articles that used deep learning and machine learning methods for brain tumor segmentation. In addition, they covered papers until mid-2021 only and did not cover studies on ViT. Similarly, the survey by Akinyelu et al [22] has a broad scope, as it covered different methods including CNNs, capsule networks, and ViT used for brain tumor segmentation. In addition, it included only 5 studies on ViT, of which 4 were from 2022. Reviews by Parvaiz et al [20], He et al [23], and Shamshad et al [25] covered the applications of ViT in medical imaging; however, the scope of all these reviews is broad, as they included different medical imaging applications. In addition, they conducted a descriptive study of ViT for various medical imaging modalities. Similarly, many relevant recent studies on ViT-based architectures have been left out, as both the reviews [20,25] were released in early 2022. Nevertheless, the aforementioned reviews could be of interest to the readers. Table 1 compares our review with the previously published review articles.

Compared with other existing reviews on ViTs and medical imaging, our study is specific to brain cancer applications and covers the most recent developments. This review provides quantitative insights into the computational complexity and the required computational resources to implement ViT architectures for brain cancer imaging. Such insights will be helpful for the researchers to choose hardware resources and graphics processing units (GPUs). This review identifies the research challenges that are specific to ViT-based approaches in brain cancer imaging applications. These discussions will raise awareness for the related research directions. This review identifies the available public data sets and highlights the need for additional data to motivate the community to develop more publicly available data sets for brain cancer research. Furthermore, this review follows a narrative synthesis approach that would help the readers follow the text quickly.

Table 1. Comparison with similar review articles.

Review title	Month and year	Scope and coverage	Comparison with our review
Vision transformers in Medical Computer Vision—A Contemplative Retrospection [20]	March 2022	• The title is specific to ViT[a]; however, the full text has a very broad scope with discussions on deep learning, CNNs[b], and ViT. • It covers different applications in medical computer vision, including the classification of disease, segmentation of tissues, registration tasks in medical images, and image-to-text applications. • It does not provide much text on brain cancer applications of ViT. • Many recent studies of 2022 are left out as the preprint was released in March 2022. • It does not provide a comparative study on the computational complexity of ViT-based models.	• Our review is also specific to ViT. • Our review is specific to brain cancer applications. • Our review includes more recent studies on ViT. • Our review provides a comparative study of the computational complexity of the ViT-based models.
Transformers in medical imaging: A survey [25]	January 2022	• It is specific to ViT. • It has a broad scope as different medical imaging applications are included. • It does not include many recent studies on ViT for brain cancer imaging (as the preprint was released in January 2022).	• Our review is also specific to ViT. • Our review is specific to brain cancer applications. • Our review includes more recent studies on ViT.
Transformers in Medical Image Analysis: A Review [23]	August 2022	• It is specific to ViT. • It has broad scope as different medical imaging applications are included. • It provides a descriptive review of ViT techniques for different medical imaging modalities. • It does not provide a quantitative analysis of the computational complexity of ViT-based methods.	• Our review is also specific to ViT. • Our review is specific to brain cancer applications. • Our review provides a comparative study of the computational complexity of the ViT-based models.
Brain Tumor Diagnosis Using Machine Learning, Convolutional Neural Networks, Capsule Neural Networks and Vision Transformers, Applied to MRI[c]: A Survey [22]	July 2022	• It covers applications specific to brain tumor segmentation. • It has a broad scope, as it includes studies on CNNs, capsule networks, and ViT. • It includes only 5 studies on ViT. • Many recent studies are left out as it covers only 4 studies from 2022. • It provides no quantitative analysis of computational complexity.	• Our review is also specific to brain cancer and brain tumor. • Our review covers more recent studies. • Our review includes 22 studies on ViT for brain cancer application. • Our review provides a comparative study of the computational complexity of the ViT-based models.
A survey of brain tumor segmentation and classification algorithms [24]	September 2021	• It has a very broad scope as it covers traditional machine learning and deep learning methods. • It covers studies until early 2021 only.	• Our review is specific to ViT. • Our review covers more recent studies.
Deep learning for brain tumor segmentation: a survey of state-of-the-art [21]	January 2021	• It has a broad scope as it covers different deep learning methods. • Many recent studies are left out.	• Our review is specific to ViT. • Our review covers more recent studies.

[a]ViT: vision transformer.

[b]CNN: convolutional neural network.

[c]MRI: magnetic resonance imaging.

Research Problem

The popularity of transformer-based approaches for medical imaging has been increasing. Many recent studies have developed new transformer-based methods for brain cancer application. Hence, there is a need to review the recent studies on how transformer-based approaches have contributed to brain cancer diagnosis, grading, and tumor segmentation. In this study,

we present a review of the advancements in ViTs for brain cancer imaging applications. We present the recent ViT architectures for brain cancer diagnosis and classification, identify the key pipelines for combining ViT with CNNs, and highlight the key challenges and issues in developing ViT-based AI techniques for brain cancer imaging. More specifically, this review aims to identify the common techniques that were developed to use ViT for brain tumor segmentation and whether ViTs were effective in enhancing the segmentation performance. This review also identifies the common modality of brain imaging data used for training ViT for brain tumor segmentation. Moreover, this review identifies the commonly used data sets for the brain tumor that contributed to developing ViT-based models. Finally, this review presents the key challenges that the researchers faced in developing ViT-based approaches for brain tumor segmentation. We believe that this review will help researchers in deep learning and medical imaging interdisciplinary fields to understand the recent developments on the topic. Furthermore, it will appeal students and researchers interested to know about the advancements in brain cancer imaging.

Methods

Overview

We performed a literature search in famous scientific databases and conducted a scoping review following the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guidelines [26]. Multimedia Appendix 1 provides the PRISMA-ScR checklist. The literature search and the study selection were performed using the steps described in the following subsections.

Search Strategy

Search Sources

We searched for relevant literature in 4 databases: PubMed, Scopus, IEEE Xplore, and Google Scholar. The search was performed between July 31 and August 1, 2022. For Google Scholar, we retained the first 300 results, as the results beyond 300 lacked relevance to the topic of this review. We also screened the reference lists of the included studies to retrieve any additional studies that fulfilled the inclusion criteria.

Search Terms

We defined the key terms for the search by referring to the available literature and by a discussion with domain experts. The search terms comprised the terms corresponding to the intervention (ie, transformers) and the target application (ie, cancer and tumor). The search strings are provided in Multimedia Appendix 2.

Search Eligibility Criteria

Our search focused on studies that reported developing ViT-based architectures for brain tumor segmentation, brain cancer diagnosis, or prognosis. We considered studies conducted between January 2017 and July 2022. We included studies that used ViT with or without combining other deep learning architectures, such as CNN, and excluded studies that used only CNN. We excluded studies that reported the diagnosis of other cancer types, such as lung cancer or colorectal cancer, and did not report the use of the model for any form of brain cancer. We included studies that used any type of brain cancer data, including brain magnetic resonance imaging (MRI) and histopathology image data. We included studies published as peer-reviewed articles or conference proceedings and excluded nonpeer-reviewed articles (preprints), short notes, editorial reviews, abstracts, and letters to the editor. We excluded survey and review articles. We did not impose any additional restrictions on the country of publication and the performance or accuracy of the ViT used in the studies. Finally, for practical reasons, we included studies published only in English.

Study Selection

Two reviewers, HA and RQ, independently screened the titles and abstracts of the studies retrieved in the search process. In abstract screening, the reviewers excluded the studies that did not fulfill the inclusion criteria. The studies retained after the title and abstract were included for full-text reading. At this stage, disagreements between the 2 reviewers (HA and RQ) were analyzed and resolved through mutual discussion. Finally, the study selection was verified by a third reviewer.

Data Extraction

We designed a custom-built data extraction sheet. Multimedia Appendix 3 presents the different fields of information in the data extraction sheet. Initially, we pilot-tested the fields in the extraction sheet by extracting data from 7 relevant studies. Two reviewers (HA and RQ) extracted the data from the included studies. The critical information extracted was the application of ViT, the architectures of ViT, the complexity of the architectures used, the pipeline for combining ViT and CNNs, the data sets and their relevant features, and the open research questions identified in the studies. The 2 reviewers resolved disagreements through mutual discussions and revisiting the full text of the relevant study where needed.

Data Synthesis

We followed a narrative approach to synthesize the data after data extraction. We categorized the included studies based on applications, such as tumor segmentation, grading, or prognosis. We also organized the studies based on data type, such as public versus private data and 2D versus 3D data. We also identified the modality of the data used in the included studies, such as MRI or pathology images. Next, we identified the most frequently used architectures of ViT and the key pipelines for incorporating ViT in cascade or parallel connections with CNN models. We also classified the included studies based on the metrics used to evaluate the performances. Finally, if available, we identified the public code repositories for the model implementation as reported in the included studies.

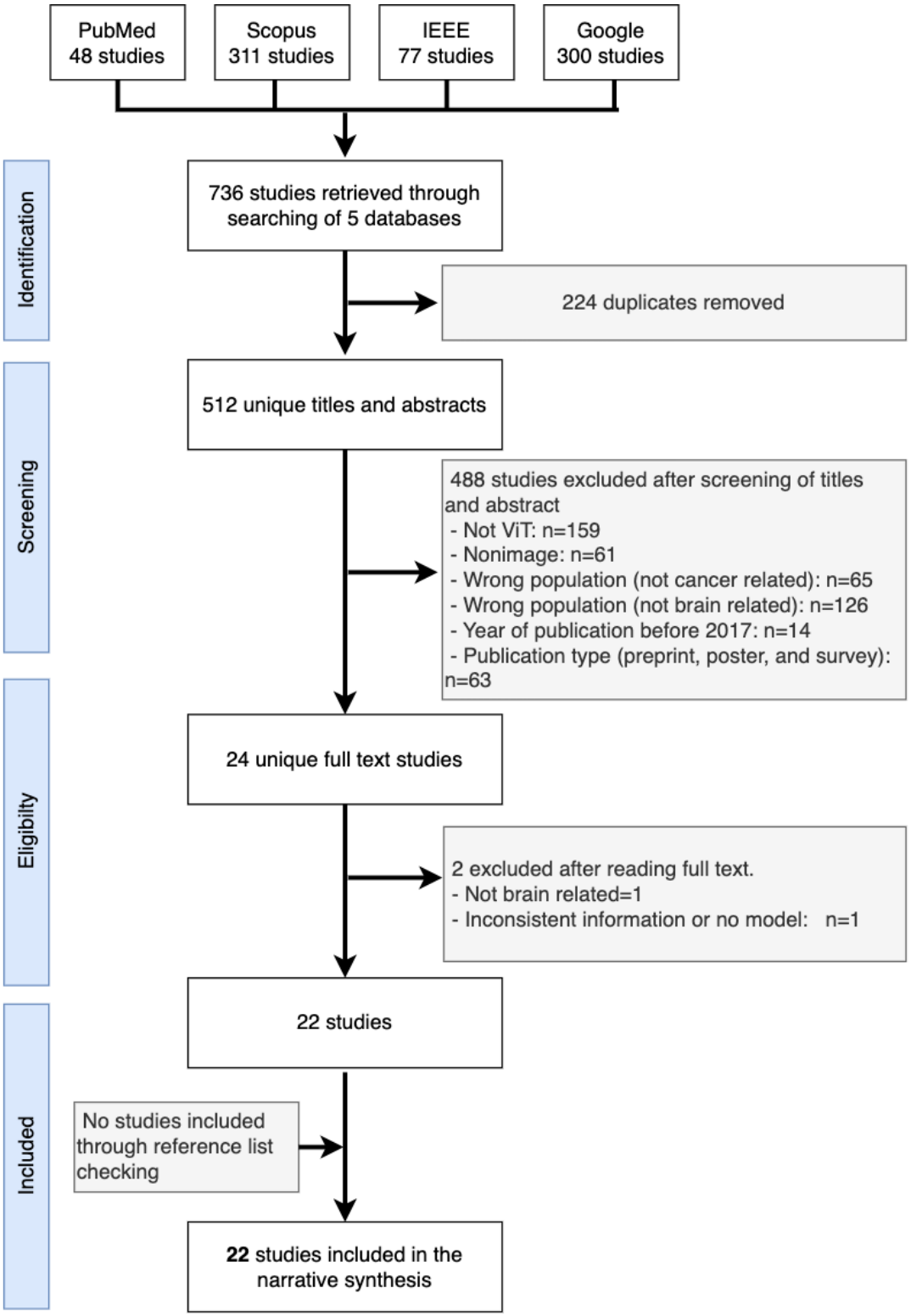
Results

Search Results

A total of 736 studies were retrieved. Of these, we removed 224 duplicates. After the title, abstract, and metadata screening, we removed 488 studies that did not fulfill the inclusion criteria and retained 24 studies. In the full-text screening, we removed

2 studies. Overall, 22 studies were included in this review. We did not find any additional studies by forward and backward reference checking. Figure 1 shows the flowchart for the study selection process. Multimedia Appendix 4 shows a list of all the included studies.

Figure 1. The PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) flowchart for the selection of the included studies. ViT: vision transformer.



Demographics of the Included Studies

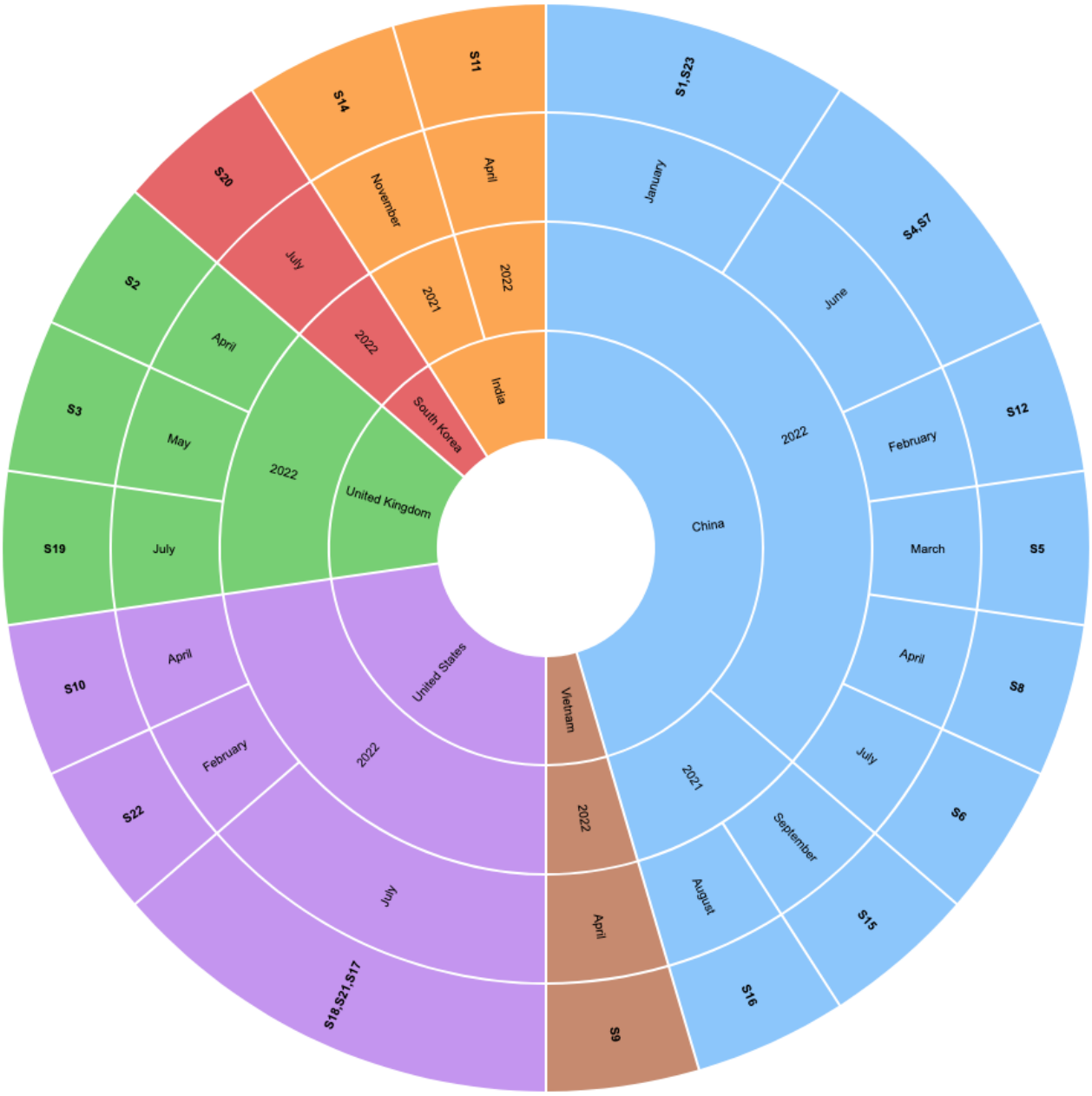
Among the 22 included studies, 9 (41%) were published in peer-reviewed journals, whereas 13 (59%) were published as conference or workshop proceedings. Of the 22 studies, 19 (86%) were published in 2022, whereas only 3 (14%) were published in 2021. No studies published before 2021 were found. Among the studies published in 2022, one-third (6/22, 27%) were published in July. The included studies were published by authors from 6 different countries (based on first-author

affiliation). Among the 22 studies, almost half (n=10, 45%) were published by authors from China and 5 (23%) were published by authors from the United States. Authors from the United Kingdom and India published 3 and 2 studies, respectively, whereas both South Korea and Vietnam published 1 study each. Multimedia Appendix 5 shows a summary of the year-wise and month-wise studies. Multimedia Appendix 6 shows a summary of the country-wise demographics of the included studies. Table 2 summarizes the demographics of the included studies. Figure 2 shows a visualization for the mapping of the included studies with year, month, and country of publication.

Table 2. Demographics of the included studies (N=22).

	Studies, n (%)
Year and month	
2022	
January	2 (9)
February	2 (9)
March	1 (4.5)
April	5 (23)
May	1 (4.5)
June	2 (9)
July	6 (27)
2021	
August	1 (4.5)
September	1 (4.5)
November	1 (4.5)
Countries	
China	10 (45)
United States	5 (23)
United Kingdom	3 (14)
India	2 (9)
South Korea	1 (4.5)
Vietnam	1 (4.5)
Type of publication	
Conference	13 (59)
Journal	9 (41)

Figure 2. Mapping of the included studies with year, month, and country. S1 through S22 are the included studies.



Main Tasks Addressed in the Studies

Among the included studies, 19 (86%) of the 22 studies addressed the task of segmentation [27-45], and 1 study [46] reported survival prediction. One study [47] reported the detection of lesions. One study [48] performed grading of the tumor. In addition, 1 study [43] performed the diagnosis of multiple sclerosis, and 1 study [45] performed reconstruction of fast MRI. One study [44] also performed isocitrate dehydrogenase (IDH) genotyping in addition to segmentation. Table 3 shows a summary of the key characteristics and tasks addressed in the included studies.

Table 3. Summary of key characteristics of the included studies.

Reference	Year	3D model	2D model	Image modality	Purpose	Transformer name	Data source
[27]	2022	Yes	Yes	MRI[a]	Segmentation	SWIN[b] transformer	Public
[28]	2022	Yes	No	MRI	Segmentation	SWIN transformer	Public
[29]	2022	Yes	No	MRI	Segmentation	SWIN transformer	Public
[30]	2022	Yes	No	MRI	Segmentation	Not available	Public
[31]	2022	Yes	No	MRI	Segmentation	Segtransvae	Public
[32]	2021	Yes	No	MRI	Segmentation	TransBTS	Public
[33]	2021	Yes	Yes	MRI	Segmentation	SegTran	Public
[34]	2022	Yes	No	MRI	Segmentation	SWIN transformer	Public
[35]	2022	Yes	No	MRI	Segmentation	TransUNet	Public
[36]	2022	Yes	No	MRI	Segmentation	Not available	Public
[37]	2022	Yes	No	MRI	Segmentation	TransBTS	Public
[38]	2022	Yes	No	MRI	Segmentation	UNETR[c]	Public
[39]	2022	Yes	No	MRI	Segmentation	SWIN transformer	Public
[40]	2021	Yes	No	MRI	Segmentation	Not available	Public
[41]	2022	No	Yes	MRI	Segmentation	Not available	Public
[42]	2022	No	Yes	MRI	Segmentation	Not available	Public+private
[43]	2022	Yes	Yes	MRI	Segmentation and diagnosis	Autoregressive transformer	Public
[44]	2022	Yes	No	MRI	Segmentation and grading	Not available	Public
[45]	2022	No	Yes	MRI	Segmentation and reconstruction	SWIN transformer	Public
[46]	2022	No	Yes	MRI	SP[d]	Not available	Public
[47]	2022	No	Yes	MRI	Detection	Not available	Private
[48]	2022	No	Yes	Pathology	Grading	Not available	Private

[a]MRI: magnetic resonance imaging.

[b]SWIN: Shifted Window.

[c]UNETR: UNet Transformer.

[d]SP: survival prediction.

Key Architectures of the ViT for Brain Tumor Segmentation

In the included studies, ViTs were combined with different variants of a CNN to improve the overall performance of brain tumor segmentation. Shifted Window (SWIN) transformer [49] has recently become a popular choice for image-based classification tasks. Therefore, the most recent studies [27-29,34,39,45] reported using SWIN transformers in their models. Some of the studies [28,29,36,38,40,41] incorporated the transformers module within the encoder or decoder or both modules of the UNet-like architectures. Some studies [30-33,35,37,44] used the transformer module as a bottleneck between the encoder and decoder modules of UNet-like architectures. One study [41] explored both cascade and parallel combinations of the transformer module with CNNs. One study [48] used the transformer module in parallel combination with a residual network (a CNN). One study [42] implemented the training of transformers using federated learning over distributed data for 22 institutions.

Complexity of the Models Used in the Studies

The included studies presented transformer-based models with different computational complexity. Of these, Fidon et al [35] used the TransUNet model, which has 116.7 million parameters, whereas the UNETR model proposed by Hatamizadeh et al [38] has 92.58 million parameters. The SegTran model proposed by Li et al [33] has 93.1 million parameters. Compared with the UNETR [38], the recent variant, that is, SWIN UNETR [34], has 61.98 million parameters. The Segtransvae [31] has 44.7 million parameters. The BTSWIN-UNet model [28] has 35.6 million parameters that are higher than other SWIN transformer–based models but much smaller than the UNETR. For example, the SWIN transformer–based models Trans-BTS and SWIN-UNet have 30.6 million and 27.1 million parameters, respectively, on the same data, but UNETR has 102.8 million

parameters on the same data. The TransConver proposed by Liang et al [27] has 9 million parameters. The SWINMR [45] has 11.40 million parameters for reconstruction. Other studies [28,30,32,36,37,39-44,46-48] did not provide details regarding the computational complexity of the models. Some studies have reported a different number of parameters for other models used on their data. We believe that these minor differences occur because of the resolution of the input images, which may not be the same in different studies.

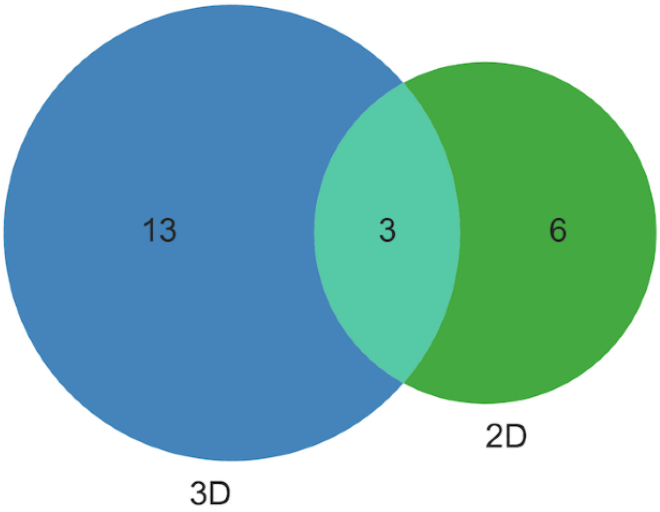
Hardware Use

Wang et al [32] used 8 NVIDIA Titan RTX GPUs for training their model. Similarly, Hatamizadeh et al [34] and Hatamizadeh et al [38] trained their models on a DGX-1 cluster with 8 NVIDIA V100 GPUs. Jia and Shu [37] used 4 NVIDIA RTX 8000 GPUs for training the model, whereas Zhou et al [48] used 4 GeForce RTX 2080 Ti GPUs. Liang et al [27] and Liang et al [29] trained their models on 2 parallel NVIDIA GeForce 2080Ti GPUs. Similarly, Huang et al [45] trained the model on 2 NVIDIA RTX 3090 GPUs with 24 GB GPU memory, and Cheng et al [44] used 2 NVIDIA V100 GPUs. Zhang et al [30] and Li et al [47] trained their models on a single NVIDIA Tesla V100 GPU, Li et al [33] trained the model on a single 24 GB Titan RTX GPU, Luu and Park [36] used a single NVIDIA RTX 3090 GPU for training the model, Liu et al [39] trained the model using NVIDIA GTX 3080, and Dhamija et al [41] used Tesla P-100 GPU.

Types of Data Used in the Studies

All the included studies (except 1 [48]) used MRI data for brain tumor segmentation. Zhou et al [48] used histopathology images. In 16 studies, volumetric MRI data were used, whereas in 9 studies, the models were developed for 2D image data. Three studies [27,33,43] reported experiments on both volumetric data and image data. Figure 3 shows the Venn diagram for the number of studies using 3D versus 2D data.

Figure 3. Venn diagrams showing the number of studies that used 3D versus 2D data.



Data Sets Used in the Studies

Three studies [42,47,48] reported using privately developed data sets or did not provide public access to the data. One study [42] used both publicly available and privately developed data. The Brain Tumor Segmentation (BraTS) challenge data set of brain MRI has been the most popular data used in 17 (77%) of the 22 studies. More specifically, 6 studies used BraTS 2021 data [28,31,34-37], 5 used BraTS 2020 data [28,32,42,44,46], 7 used BraTS 2019 data [27-29,32,33,39,40], 3 used BraTS 2018 data [27,29,43], and 1 used BraTS 2017 data [45]. Some of these studies also used >1 data set, either independently or by combining them. Other data used in the included studies were MRI data from the Medical Decathlon used by Hatamizadeh et al [38], the Cancer Imaging Archive data used by Dhamija at [41], the UK Biobank data used by Pinaya et al [43], data from the University Hospital of Ljubljana used by Pinaya et al [43], the Calgary-Campinas Magnetic Resonance reconstruction data used by Huang et al [45], data from the University Hospital of Patras Greece used by Zhou et al [48], and data from the Cancer Hospital and Shenzhen Hospital used by Li et al [47]. One study [30] did not specify the data. Table 4 summarizes the data sets used in the included studies and provides the public access links for each data set. Figure 4 shows the Venn diagram for the number of studies using public versus private data.

Table 4. Data sets used in the included studies.

Data set name	Modality	Available	URL	Used by the following studies
BraTS[a] 2021	MRI[b]	Public	[50]	[28,31,34-37]
BraTS 2020	MRI	Public	[51]	[28,32,42,44,46]
BraTS 2019	MRI	Public	[52]	[27-29,32,33,39,40]
BraTS 2018	MRI	Public	[53]	[27,29,43]
BraTS 2017	MRI	Public	[50]	[45]
Decathlon	MRI	Public	[54]	[38]
TCIA[c]	MRI	Public	[55]	[41]
UK Biobank	MRI	Public	[56]	[43]
University Hospital of Ljubljana	MRI	Public	[57]	[43]
Calgary-Campinas MR[d] reconstruction data set	MRI	Public	[58]	[45]
University Hospital of Patras Greece	Pathology images	Private	—[e]	[48]
Cancer Hospital and Shenzhen Hospital data	—	Private	—	[47]
Not specified	N/A[f]	N/A	N/A	[30,47]

[a]BraTS: brain tumor segmentation.

[b]MRI: magnetic resonance imaging.

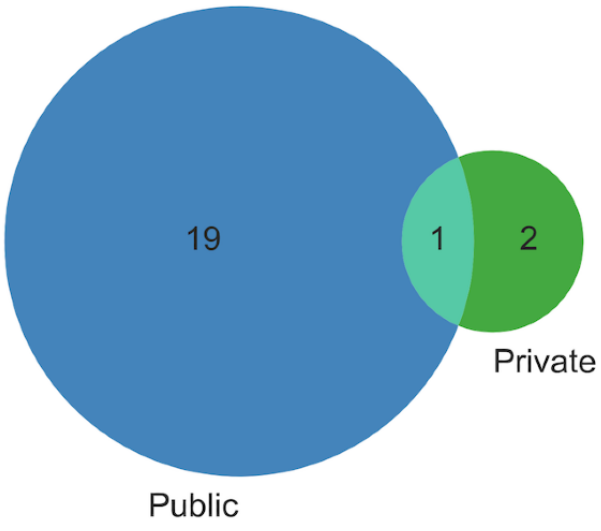
[c]TCIA: The Cancer Imaging Archive.

[d]MR: magnetic resonance.

[e]Not available.

[f]N/A: not applicable.

Figure 4. Venn diagrams showing the number of studies that used public versus private data sets.



Evaluation Metrics

The Dice score and the Hausdorff distance measurements are popular metrics commonly used to evaluate segmentation performance on the BraTS MRI data sets. Hence, in the included studies, the Dice score and Hausdorff distance were the most common metrics used to assess the results of brain tumor segmentation. In summary, 19 studies [27-45] reported the use of the Dice score, whereas 15 studies [27-32,34-40,42,44] used both the Dice score and Hausdorff distance. Two studies [41,45] reported intersection-over-union. One study [42] reported the focal score and Tversky score for the federated learning framework evaluation in addition to the Dice score and Hausdorff distance for the segmentation evaluation. One study [45] reported peak signal:noise ratio, structural similarity index, and Fréchet Inception Distance in the assessment of the reconstructed MRI in addition to Intersection over Union and Dice scores for segmentation evaluation. One study [46] used

the concordance index and hazard ratio to evaluate the performance of survival analysis. One study [47] reported sensitivity and precision, and 1 study [48] reported precision and recall.

Discussion

Principal Findings

In this study, we reviewed the studies that used ViT to aid in brain cancer imaging applications. We found that most studies (19/22, 86%) were published in 2022, and almost one-third of these studies (6/19, 32%) were published in the second quarter of 2022. As ViT was first proposed in 2020 for natural images, it has only recently been explored in brain MRI and cancer imaging. Almost half of the studies (10/22, 45%) were published by authors from China. Furthermore, the authors from China published twice the number of studies published by authors from the United States. Other countries published approximately one-third of the studies (7/22, 32%).

Motivation of Using Transformers for Segmentation

The transformer module works on the self-attention concept, that is, calculating pairwise interactions between all input units. Thus, transformers are good at learning contextualized features. Although this learning of the contextualization by a transformer can be related to the upsampling path in a UNet encoder-decoder architecture, the transformer overcomes the limitation of the receptive field, and hence, it works better to capture long-range correlations [34]. In a UNet architecture, one may enlarge the receptive fields by adding more downsampling layers or by introducing larger stride sizes in the convolution operations of the downsampling path. However, the former increases the number of parameters and may lead to overfitting, whereas the latter sacrifices the spatial precision of the feature maps [34]. Nevertheless, the initial attempts to introduce transformers for brain tumor segmentation used the transformer block in the encoder or decoder or the bottleneck stage of the UNet-like architectures. These approaches were mainly driven by the success of UNet-based architectures for segmentation, such as nnUNet's success on the BraTS2020 challenge [59]. In addition, until 2020, CNN-based models were the best performers for brain tumor segmentation. Therefore, nnUNet [59] was the winning entry for the BraTS2020 challenge. With improved strategies and architectures, attention-based models performed competitively in recent years. Wang et al [32] presented the TransBTS model, which was the first attempt to incorporate transformers into a 3D CNN for brain tumor segmentation. Although Hatamizadeh et al [34] reported SWIN UNETR for brain tumor segmentation, and it was the first transformer-based model that performed competitively for the BraTS 2021 segmentation task. The TransBTS model was trained and tested on the BraTS2018 and BraTS2019 data sets, whereas the SWIN UNETR has been evaluated on the BraTS 2021 data set. However, for the BraTS 2021 data set, the winning entry was an extension of the nnUNet model [59] presented by Luu and Park [36] who proposed introducing attention in the decoder of the nnUNet to perform the tumor segmentation. As identified by Jia and Shu [37], the UNETR removed convolutional blocks in the encoder, which may result in insufficient extraction of local context information when applied to volumetric MRI data. Overall, these approaches of combining transformers and CNNs are driven by the motivation to use the best of both worlds. These studies suggested that the best-of-both-worlds approach can be effective in improving brain tumor segmentation by combining CNNs with transformers. In theory, there are many possibilities for how we approach combining the advantages offered by the 2 different architectures.

Applications Covered in the Studies

Most of the studies included are those that either designed an attention-based architecture or used existing ViT architectures to achieve the task of tumor segmentation. In the brain segmentation tasks, the key focus is the segmentation of gliomas, which is the most common brain tumor. As most of these studies used 1 of the variants of the BraTS data set where the MRI data are annotated for 4 regions, these studies reported segmentation of the whole tumor, tumor core, enhancing tumor, and background. Some studies also reported using attention-based models for other applications related to brain cancer, such as survival prediction, MRI reconstruction, grading of brain cancer, and IDH genotyping.

Discussion Related to the Architectures

Among the studies that used the ViT module after a 3D CNN features extraction, the TransBTS [32] was the first architecture (released in September 2021) and served as inspiration for many other architectures. The TransBTS architecture was motivated by the idea of incorporating global context into the volumetric spatial features of brain tumors. Furthermore, the work highlighted the need to use an attention module on image patches instead of flattened images, unlike previous efforts. Essentially, the flattening of high-resolution images makes the implementation impractical, as transformers have a quadratic computational complexity with respect to the number of tokens (ie, the dimension of the flattened image). The TranBTS architecture has downsampling and upsampling layers linked through skip connections; however, in the bottom part of the architecture, there are transformer layers that help with the global context capturing. These transformer layers are in addition to a linear projection layer and a patch embedding layer to transfer the image to sequence representation. So, in a way, the ViT serves as the bottleneck layer to capture long-range dependencies. Later, Jia and Shu [37] presented a modification in the TransBTS architecture [32] using 2 ViT blocks after the encoder part instead of 1 transformer block in the TransBTS. Specifically, the outputs of the fourth and fifth downsampling layers pass through a feature embedding of a feature representation layer, transformer layers, and a feature mapping layer and then pass through the corresponding upsampling 3D CNN layers. Compared with the TransBTS architecture, where the transformer was used at the end of the encoder and features representation was obtained after the fourth layer, Jia and Shu [37], increased the depth to 5 layers and used the transformer in both the fourth and fifth layers. Therefore, after the fourth layer, the transformer effectively builds a skip connection with the corresponding layer of the decoder block.

Similarly, Zhang et al [30] used a multihead self-attention–based transcoder module embedded after the encoder of a 3D UNet.

However, they replaced the residual blocks of the 3D UNet with a self-attention layer that operated on a 3D feature map, followed by progressive upsampling via a 3D CNN decoder module. Pham et al [31] also used transformer layers after a 3D CNN module and used a variational encoder to reconstruct the volumetric images. Li et al [33] presented the SegTran architecture, which is again based on using the transformer modules after the features extraction with CNN, thus capturing the global context. Here, the authors suggested combining the CNN features with positional encodings of the pixel coordinates and flattening them into a sequence of local feature vectors.

Fidon et al [35] used the TransUNet architecture [60] as the backbone of their model and used the test time augmentation strategy to improve inference. Finally, Cheng et al [44] presented the MTTUNet architecture, which is a UNet-like encoder-decoder architecture for multitasking. They used the CNN layers to extract spatial features, which were then processed by the bottleneck transformer block. Subsequently, the decoder network performed the segmentation task. In addition, the authors also used the transformer output to perform IDH genotyping, thus making it a multitask architecture.

Hatamizadeh et al [38] presented the UNETR architecture that redefined the task of 3D segmentation as a 1D sequence-to-sequence classification that can be used with a transformer to learn contextual information. Therefore, the transformer block in the UNETR operates on the embedded representation of the 3D MRI input data. In effect, the transformer is incorporated within the encoder part of a UNet architecture. Compared with other architectures such as BTSWIN-UNet [30], TransBTS [32], SegTran [33,35], and BiTr-UNet [37], which use the transformer as a bottleneck layer of the encoder-decoder architectures, the UNETR directly connects the encoded representation from the encoder with the decoder part. Compared with other methods where the encoder part uses 3D CNN blocks, such as TransBTS [32] and BiTr-UNet [37], the UNETR does not use a convolutional block in the encoder. Instead, the UNETR obtains a 2D representation for the 3D volumes and then uses the 2D ViT architecture that works on the 2D patches of the images. Each patch is treated as 1 token for the attention operation. UNETR does not rely on a backbone CNN for generating the input sequences and directly uses the tokenized patches.

Luu and Park [36] introduced an attention mechanism in the decoder of the nnUNet [59] to perform the tumor segmentation. They extended the nnUNet and modified it by using axial attention in the decoder of the 3D UNet. Furthermore, they doubled the number of filters in the encoder while retaining the same number in the decoder. Sagar [40] presented the Vision Transformer for Biomedical Image Segmentation architecture, which used transformer blocks in the encoder and decoder of a UNet architecture. The architecture introduced multiscale convolutions for feature extraction that were used as input to the transformer block.

Dhamija et al [41] explored the sequential and parallel stacks of transformer-based blocks using a UNet block. In principle, they used a transformer-based encoder and a CNN-based decoder connected in parallel with a UNet-based encoder and then in cascade with a UNet-based encoder. Apparently, the parallel combination (USegTransformer-P) outperformed the cascade combination by some margin. Zhou et al [48] designed a parallel dual-branch network of a CNN (the ResNet architecture) and ViT and used it to grade brain cancer from pathology images. The dual-branch network established a duplex communication between the ResNet and ViT blocks that sends global information from the ViT to ResNet and local information from ResNet to the ViT.

Many similar architectures were probably released concurrently by different research groups or released very close in time to each other. For example, Li et al [33] found that segmentation transformer [61] and TransUNet [60] were released concurrently with their own model. Therefore, it is not surprising that there are a few similarities between the approaches adopted by these studies.

Discussion Related to SWIN Transformers

In general, transformers are notoriously popular for the computational complexity of the order O (n^2). For example, as identified by Jia and Shu [37], UNETR stacks transformer layers and keeps the sequence data dimension unchanged during the entire process, which results in expensive computation for high-resolution 3D images. SWIN transformers helped overcome the computational complexity. Hence, it became a popular backbone architecture for many recent studies [27-29,32,39,45] to overcome the computational complexity of transformer-based models. For example, Liang et al [27] reported the use of a 2D SWIN transformer [49] and a 3D SWIN transformer [62] to replace the traditional architecture of ViT to overcome the computational complexity. Jiang et al [28] used a SWIN transformer as the encoder and decoder rather than as the attention layer. Furthermore, they extended the 2D SWIN transformer to a 3D variant that provided a base module. Similarly, Liang et al [29] used a 3D SWIN transformer block in the encoder and decoder of a 3D UNet-like architecture. The architecture was inspired by the SWIN transformer and the SWIN-UNet model; however, they replaced the patchify stem with a convolutional stem to stabilize the model training. Furthermore, they used overlapping patch embedding and downsampling, which helped to enhance the locality of the segmentation network.

Hatamizadeh et al [34] extended the UNETR architecture to the SWIN-UNet transformer (SWIN UNETR), which incorporated a SWIN transformer in the encoder part of the 3D UNet. The decoder part still used a CNN architecture to upsample the features to the segmentation masks. As reported previously, the SWIN UNETR was the first transformer architecture that performed competitively on the BraTS 2021 segmentation challenge. Liu et al [39] presented a transition net architecture that combined a 2D SWIN transformer with a 3D transition decoder. The transition block transforms the 3D volumetric data into a 2D representation, which is then provided as an input to the SWIN transformer. Subsequently, in the decoder part, the transition block transforms the multiscale feature maps into a 3D representation to obtain the segmentation results. Huang et al [45] used a cascade of residual SWIN transformers to build

a feature extraction module, followed by a 2D CNN network. This architecture was designed for MRI reconstruction.

Discussion Related to Model Complexity

In general, transformer architectures have a high computational complexity. The number of parameters for the architectures for the models, such as UNETR and TransUNet, are as large as 92 million and 116 million, respectively. The SWIN transformer-based architecture has a relatively smaller number of parameters (of the order of 30-45 million). For models with a higher number of parameters, the researchers had to rely on high-end GPU resources. Therefore, the computational setup reported in some of the included studies was built with as many as 8 GPUs. However, few studies also reported training the models on a single GPU with memory sizes ranging from 12 GB to 24 GB.

Discussion Related to 3D Data

Our categorization of a model designed for 3D or 2D data was either based on direct extraction of the information from the studies or the description of the model architecture in the included studies. Therefore, if a study did not specify whether it used the volumetric data directly or transformed the data into 2D images but provided a 2D model architecture, we placed the study in the 2D data category. Many modern deep learning methods for medical imaging, including transformers, rely on pretrained models as their backbones. These backbones can generalize well, making them good candidates for use in other related tasks, as they provide generalization, better convergence, and improved segmentation performance [39]. However, Liu et al [39] argued that such backbone architectures are, in general, difficult to be migrated to 3D brain tumor segmentation. First, there is a general lack of 3D data, and most publicly available data sets provide 2D data. Second, medical images such as MRI vary in their distribution and style compared with natural images. These variations hinder the direct transformation of the 2D pretrained models for 3D volumetric data. Hence, they recommended transforming the 3D data into a 2D representation to enable its use with 2D transformers. However, numerous other studies have developed and used 3D models directly on volumetric data.

The most commonly used data in the included studies were the brain MRI of the BraTS data set. The BraTS data set has been phenomenal in facilitating the research on brain glioma segmentation. The BraTS challenge has served as a dedicated venue for the last 11 years and has established itself as a foundation data set in helping the community push the state-of-the-art in brain tumor segmentation. The BraTS data set has 4 MRI modalities, namely, T1-weighted, postcontrast T1-weighted, T2-weighted, and T2 fluid-attenuated inversion recovery. Furthermore, the data set provides baseline segmentation annotation from physicians.

Discussion Related to Evaluation Metrics

The Dice score and Hausdorff distance measurements have been more commonly reported, as these metrics are widely used to evaluate segmentation performance on the BraTS MRI data sets. In the included studies, the Dice score and Hausdorff distance were the most common metrics used to assess the results of brain tumor segmentation.

Strengths and Limitations

Strengths

Although there has been a surge in studies on the use of ViTs in medical imaging, only a few reviews have been reported on ViTs in medical imaging [20,23,25]; however, their scopes are too broad. In comparison, to the best of our knowledge, this is the first review of the applications and potential of ViTs to enhance the performance of brain tumor segmentation. This review covers all the studies that used ViTs for brain cancer imaging; thus, this is the most comprehensive review. This review is helpful for the community interested in knowing the different architectures of ViTs that can help in brain tumor segmentation. Unlike other reviews [20,23,25] that cover many different medical imaging applications, this review focuses on studies that have only developed ViTs for brain tumor segmentation. In this review, we followed the PRISMA-ScR guidelines [26]. We retrieved articles from the popular web-based libraries of medical science and computing to include as many relevant studies as possible. We avoided bias in study selection through an independent selection of studies by 2 reviewers and through validation of the selected studies and data extraction by the third reviewer. This review provides a comprehensive discussion on the different pipelines to combine ViTs with CNNs. Hence, this review will be very useful for the community to learn about the different pipelines and their working for brain tumor segmentation. In addition, we identify the computational complexity of the various pipelines to help the readers understand the associated computational cost of ViTs for brain tumor segmentation. We provide a comprehensive list of available data sets for brain MRI and hope that it will provide a good reference point for researchers to identify suitable data sets for developing models for BraTS. We maintain an active web-based repository that will be populated with relevant studies in the future.

Limitations

In this review, we included studies from 4 major databases. Despite our best efforts to retrieve as many studies as possible, the possibility that some relevant studies may be missed cannot be ruled out. Moreover, the number of publications on the applications of ViTs in medical imaging is increasing at an unprecedented rate; hence, recent studies may be published while we draft this work. For practical reasons, we only included studies in English. Therefore, non-English text might be excluded even if it were relevant. Not all studies reported on the computational complexity and the required training time. Hence, we provide the computational complexity only for the studies in which this information was available; thus, the comparison might not be exhaustive. This review did not analyze the claims on the performance of the different architectures, as such an assessment is beyond the scope of this work. We did not attempt to reproduce the results reported in the studies, as such an execution of the computer code is beyond the scope of the review. We included studies that reported working with any imaging modality for brain cancer and did not evaluate the use of physiological signals, although understanding physiological

signals can also play a significant role in brain cancer studies. We did not evaluate the bias in the training data used in the included studies; therefore, the performance reported for ViTs in brain cancer imaging could be occasionally overestimated.

Open Questions and Challenges

Research efforts on developing transformer-based methods for brain cancer applications are progressing rapidly. Some of the challenges are highlighted in the following text.

In the included studies, we did not find any study that addresses the challenge of early detection of brain cancer. Similarly, the number of studies related to prognosis and tumor growth in the brain is also minimal. Early detection and prognosis are applications of great interest where the potential of ViTs can be explored. One approach is to combine ViT with the sequential representation of time-based data for tumor growth in the brain.

ViTs lack scale invariance, rotation invariance, and inductive bias capabilities. Consequently, they do not perform well at capturing local information and cannot be trained well with a small amount of data [48]. One way to overcome this limitation is to provide a larger training data set. Therefore, the development of large public data sets is encouraged. Another widely used method in the included studies is combining ViTs with CNNs.

In general, models pretrained on a large-scale data set (ImageNet) are known to perform well on many other data sets. However, using the pretrained transformer-based models and fine-tuning them for brain cancer imaging did not improve the performance, as reported by Hatamizadeh et al [38]. Similarly, Pinaya et al [43] reported that the model trained on 3D data from the UK Biobank could perform well on the test set. However, the performance degraded when the model was evaluated on subsets of other data sets. Therefore, the generalization of the models is still a challenge.

Combining CNN with ViTs can be achieved through serial (cascade), parallel connections, or a combination of both. In serial combination of CNNs and ViTs, the arrangement may cause training ambiguities in terms of fusing local and global features. If the learning eventually loses local and global dependencies in the image data [48,63,64], optimal performance may not be achieved. In contrast, for parallel combinations, there will be undesired redundant information captured by the 2 models [33].

The BraTS challenge completed its 10 years in 2021 and has been a dedicated venue for facilitating the state-of-the-art developments of methods for glioma segmentation [37]. As the data set is publicly available, almost all the included studies have used it. However, there seems to be a very limited effort in developing other data sets that are publicly available. It would be interesting to have additional data sets for brain cancer imaging that can facilitate advancing the research on AI models for brain cancer diagnosis and prognosis.

The included studies reported advancements in transformer-based architectures for brain cancer imaging. However, these studies commonly lack the explanability and interpretability of the model behavior. Future research should focus on new methods to address this issue.

ViT-based architectures, as of now, may not always be the best for brain tumor segmentation. For example, the TransBTS model (a ViT-based model) had suboptimal performance owing to its inherently inefficient architecture, where the ViT is only used in the bottleneck as a stand-alone attention module and does not have a connection to the decoder at different scales (as identified by Hatamizadeh et al [34]). In contrast, architectures based on UNet (eg, nnUNet and SegResNet) have achieved competitive benchmarks on the BraTS challenge.

As identified by Huang et al [45], one can argue that the heavy computations in transformers are the main bottleneck in development, and the performance improvements of transformers for brain cancer imaging come at the cost of computational complexity. Therefore, lightweight implementations of transformer architectures for brain cancer imaging are a topic of great interest for future research. Furthermore, the transformer architectures that transform image data into sequential representation (such as in UNETR) may not be the best choice. First, the removal of convolutional blocks in the encoder does not guarantee the capture of context information in volumetric MRI data. Second, keeping a fixed sequence during the entire processing of data leads to expensive computation when the input data are a batch of high-resolution 3D images [37]. Models such as UNETR and TransBTS for brain tumor segmentation lack cross-plane contextual information; hence, the 3D spatial context is not fully captured by these models [29].

Conclusions

In this work, we performed a scoping review of 22 studies that reported ViT-based AI models for brain cancer imaging. We identified the key applications of ViTs in developing AI models for tumor segmentation and grading. ViTs have enabled researchers to push the state-of-the-art in brain tumor segmentation, although such an improvement has resulted in a trade-off between model complexity and performance. We also summarized the different vision architectures and the pipelines with ViTs as the backbone architecture. We also identified the commonly used data sets brain tumor segmentation tasks. Finally, we provided insights into the key challenges in advancing brain cancer diagnosis or prognosis using ViT-based architectures. Although ViT-based architectures have great potential in advancing AI methods for brain cancer, clinical transformations can be challenging, as these models are computationally complex and have limited or no explainability. We believe that the findings of this review will be beneficial to the researchers studying AI and cancer.

Authors' Contributions

HA contributed to the conception, design, literature search, data selection, data synthesis, data extraction, and drafting. RQ contributed to the data synthesis, data extraction, and drafting. ZS contributed to the drafting and critical revision of the manuscript. All authors gave their final approval and accepted accountability for all aspects of this work.

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist.
[DOCX File , 108 KB - medinform_v11i1e47445_app1.docx]

Multimedia Appendix 2
Search terms.
[DOCX File , 20 KB - medinform_v11i1e47445_app2.docx]

Multimedia Appendix 3
Extraction fields.
[DOCX File , 22 KB - medinform_v11i1e47445_app3.docx]

Multimedia Appendix 4
Included studies.
[XLSX File (Microsoft Excel File), 41 KB - medinform_v11i1e47445_app4.xlsx]

Multimedia Appendix 5
Demographics of the included studies showing month-wise publications.
[PNG File , 49 KB - medinform_v11i1e47445_app5.png]

Multimedia Appendix 6
Demographics of the included studies showing country-wise publications.
[PNG File , 99 KB - medinform_v11i1e47445_app6.png]

References

1. Koo YL, Reddy GR, Bhojani M, Schneider R, Philbert MA, Rehemtulla A, et al. Brain cancer diagnosis and therapy with nanoplatforms. Adv Drug Deliv Rev 2006 Dec 01;58(14):1556-1577. [doi: 10.1016/j.addr.2006.09.012] [Medline: 17107738]
2. Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ. Multimodal biomedical AI. Nat Med 2022 Sep;28(9):1773-1784. [doi: 10.1038/s41591-022-01981-2] [Medline: 36109635]
3. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med 2022 Jan;28(1):31-38. [doi: 10.1038/s41591-021-01614-0] [Medline: 35058619]
4. Saporta A, Gui X, Agrawal A, Pareek A, Truong SQ, Nguyen CD, et al. Benchmarking saliency methods for chest X-ray interpretation. Nat Mach Intell 2022 Oct 10;4(10):867-878 [FREE Full text] [doi: 10.1038/s42256-022-00536-x]
5. Mohsen F, Ali H, El Hajj N, Shah Z. Artificial intelligence-based methods for fusion of electronic health records and imaging data. Sci Rep 2022 Oct 26;12(1):17981 [FREE Full text] [doi: 10.1038/s41598-022-22514-4] [Medline: 36289266]
6. Ali H, Biswas MR, Mohsen F, Shah U, Alamgir A, Mousa O, et al. The role of generative adversarial networks in brain MRI: a scoping review. Insights Imaging 2022 Jun 04;13(1):98 [FREE Full text] [doi: 10.1186/s13244-022-01237-0] [Medline: 35662369]
7. Huang H, Lin L, Tong R, Hu H, Zhang Q, Iwamoto Y, et al. UNet 3+: a full-scale connected UNet for medical image segmentation. In: Proceedings of the 2020 International Conference on Acoustics, Speech and Signal Processing. 2020 Presented at: ICASSP '20; May 4-8, 2020; Barcelona, Spain p. 1055-1059 URL: https://ieeexplore.ieee.org/document/9053405 [doi: 10.1109/icassp40776.2020.9053405]
8. Mubashar M, Ali H, Grönlund C, Azmat S. R2U++: a multiscale recurrent residual U-Net with dense skip connections for medical image segmentation. Neural Comput Appl 2022;34(20):17723-17739 [FREE Full text] [doi: 10.1007/s00521-022-07419-7] [Medline: 35694048]
9. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Proceedings of the 31st Conference on Neural Information Processing Systems. 2017 Presented at: NIPS '17; December 4-9, 2017; Long Beach, CA p. 1-11 URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

10. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. Transformers: state-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. 2020 Presented at: EMNLP '20; November 16-20, 2020; Virtual Event p. 38-45 URL: https://aclanthology.org/2020.emnlp-demos.6.pdf [doi: 10.18653/v1/2020.emnlp-demos]
11. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: transformers for image recognition at scale. arXiv. Preprint posted online June 3, 2021 2021 [FREE Full text] [doi: 10.48550/arXiv.2010.11929]
12. Gao X, Khan MH, Hui R, Tian Z, Qian Y, Gao A, et al. COVID-VIT: classification of Covid-19 from 3D CT chest images based on vision transformer model. In: Proceedings of the 3rd International Conference on Next Generation Computing Applications. 2022 Presented at: NextComp '22; October 6-8, 2022; Flic-en-Flac, Mauritius p. 1-4 URL: https://ieeexplore.ieee.org/document/9932246 [doi: 10.1109/nextcomp55567.2022.9932246]
13. Zhang L, Wen Y. A transformer-based framework for automatic COVID19 diagnosis in chest CTs. In: Proceeedings of the 2021 International Conference on Computer Vision Workshops. 2021 Presented at: ICCVW '21; October 11-17, 2021; Montreal, BC p. 513-518 URL: https://ieeexplore.ieee.org/document/9607582 [doi: 10.1109/iccvw54120.2021.00063]
14. Costa GS, Paiva AC, Júnior GB, Ferreira MM. COVID-19 automatic diagnosis with CT images using the novel transformer architecture. Simpósio Brasileiro de Computação Aplicada à Saúde 2021:293-301 [FREE Full text] [doi: 10.5753/sbcas.2021.16073]
15. Marchiori E, Tong Y, van Tulder G. Multi-view analysis of unregistered medical images using cross-view transformers. In: Proceedings of the 24th International Conference on Medical Image Computing and Computer Assisted Intervention. 2021 Presented at: MICCAI '21; September 27-October 1, 2021; Strasbourg, France p. 104-113 URL: https://dl.acm.org/doi/abs/10.1007/978-3-030-87199-4_10 [doi: 10.1007/978-3-030-87199-4_10]
16. Zhang Z, Zhang W. Pyramid medical transformer for medical image segmentation. arXiv. Preprint posted online April 29, 2021 2021 [FREE Full text]
17. Xie Y, Zhang J, Shen C, Xia Y. CoTr: efficiently bridging CNN and transformer for 3D medical image segmentation. In: Proceedings of the 24th International Conference on Medical Image Computing and Computer Assisted Intervention. 2021 Presented at: MICCAI '21; September 27-October 1, 2021; Strasbourg, France p. 171-180 URL: https://link.springer.com/chapter/10.1007/978-3-030-87199-4_16 [doi: 10.1007/978-3-030-87199-4_16]
18. He S, Grant PE, Ou Y. Global-local transformer for brain age estimation. IEEE Trans Med Imaging 2022 Jan;41(1):213-224 [FREE Full text] [doi: 10.1109/TMI.2021.3108910] [Medline: 34460370]
19. Chen RJ, Chen C, Li Y, Chen TY, Triser AD, Krishnan RG, et al. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022 Presented at: CVPR '22; June 19-24, 2022; New Orleans, LA p. 16123-16134 URL: https://ieeexplore.ieee.org/document/9880275/ [doi: 10.1109/cvpr52688.2022.01567]
20. Parvaiz A, Khalid MA, Zafar R, Ameer H, Ali M, Fraz MM. Vision Transformers in medical computer vision—a contemplative retrospection. Eng Appl Artif Intell 2023 Jun;122:106126 [FREE Full text] [doi: 10.1016/j.engappai.2023.106126]
21. Magadza T, Viriri S. Deep learning for brain tumor segmentation: a survey of state-of-the-art. J Imaging 2021 Jan 29;7(2):19 [FREE Full text] [doi: 10.3390/jimaging7020019] [Medline: 34460618]
22. Akinyelu AA, Zaccagna F, Grist JT, Castelli M, Rundo L. Brain tumor diagnosis using machine learning, convolutional neural networks, capsule neural networks and vision transformers, applied to MRI: a survey. J Imaging 2022 Jul 22;8(8):205 [FREE Full text] [doi: 10.3390/jimaging8080205] [Medline: 35893083]
23. He K, Gan C, Li Z, Rekik I, Yin Z, Ji W, et al. Transformers in medical image analysis: a review. Intell Med 2023 Feb;3(1):59-78 [FREE Full text] [doi: 10.1016/j.imed.2022.07.002]
24. Biratu ES, Schwenker F, Ayano YM, Debelee TG. A survey of brain tumor segmentation and classification algorithms. J Imaging 2021 Sep 06;7(9):179 [FREE Full text] [doi: 10.3390/jimaging7090179] [Medline: 34564105]
25. Shamshad F, Khan S, Zamir SW, Khan MH, Hayat M, Khan FS, et al. Transformers in medical imaging: a survey. Med Image Anal 2023 Aug;88:102802. [doi: 10.1016/j.media.2023.102802] [Medline: 37315483]
26. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
27. Liang J, Yang C, Zeng M, Wang X. TransConver: transformer and convolution parallel network for developing automatic brain tumor segmentation in MRI images. Quant Imaging Med Surg 2022 Apr;12(4):2397-2415 [FREE Full text] [doi: 10.21037/qims-21-919] [Medline: 35371952]
28. Jiang Y, Zhang Y, Lin X, Dong J, Cheng T, Liang J. SwinBTS: a method for 3D multimodal brain tumor segmentation using Swin transformer. Brain Sci 2022 Jun 17;12(6):797 [FREE Full text] [doi: 10.3390/brainsci12060797] [Medline: 35741682]
29. Liang J, Yang C, Zhong J, Ye X. BTSwin-Unet: 3D U-shaped symmetrical Swin transformer-based network for brain tumor segmentation with self-supervised pre-training. Neural Process Lett 2022 Jun 17;55(4):3695-3713 [FREE Full text] [doi: 10.1007/s11063-022-10919-1]

30. Zhang T, Xu D, He K, Zhang H, Fu Y. 3D U-Net with trans-coder for brain tumor segmentation. In: Proceedings of the 13th International Conference on Graphics and Image Processing. 2021 Presented at: ICGIP '21; February 16, 2022; Kunming, China URL: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/12083/120831Q/3D-U-Net-with-trans-coder-for-brain-tumor-segmentation/10.1117/12.2623549.short [doi: 10.1117/12.2623549]
31. Pham QD, Nguyen-Truong H, Phuong NN, Nguyen KN, Nguyen CD, Bui T, et al. Segtransvae: hybrid Cnn - transformer with regularization for medical image segmentation. In: Proceedings of the 19th International Symposium on Biomedical Imaging. 2022 Presented at: ISBI '22; March 28-31, 2022; Kolkata, India p. 1-5 URL: https://ieeexplore.ieee.org/document/9761417 [doi: 10.1109/isbi52829.2022.9761417]
32. Wang W, Chen C, Ding M, Yu H, Zha S, Li J. TransBTS: multimodal brain tumor segmentation using transformer. In: Proceedings of the 24th International Conference on Medical Image Computing and Computer Assisted Intervention. 2021 Presented at: MICCAI '21; September 27-October 1, 2021; Strasbourg, France p. 109-119 URL: https://link.springer.com/chapter/10.1007/978-3-030-87193-2_11 [doi: 10.1007/978-3-030-87193-2_11]
33. Li S, Sui X, Lou X, Xu X, Liu Y, Goh R. Medical image segmentation using squeeze-and-expansion transformers. In: Proceedings of the 30eth International Joint Conference on Artificial Intelligence. 2021 Presented at: IJCAI '21; August 19-26, 2021; Virtual Event p. 807-815 URL: https://www.ijcai.org/proceedings/2021/0112.pdf [doi: 10.24963/ijcai.2021/112]
34. Hatamizadeh A, Nath V, Tang Y, Yang D, Roth HR, Xu D. Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images. In: Proceedings of the Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 7th International Workshop. 2021 Presented at: BrainLes '21; September 27, 2021; Virtual Event p. 284 URL: https://dl.acm.org/doi/abs/10.1007/978-3-031-08999-2_22 [doi: 10.1007/978-3-031-08999-2_22]
35. Fidon L, Shit S, Ezhov I, Peetzold JC, Ourselin S, Vercauteren T. Generalized wasserstein dice loss, test-time augmentation, and transformers for the BraTS 2021 challenge. In: Proceedings of the 7th International Workshop on Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. 2022 Presented at: BrainLes '21; September 27, 2021; Virtual Event p. 187-196 URL: https://link.springer.com/chapter/10.1007/978-3-031-09002-8_17 [doi: 10.1007/978-3-031-09002-8_17]
36. Luu HM, Park SH. Extending nn-UNet for brain tumor segmentation. In: Proceedings of the Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 7th International Workshop. 2021 Presented at: BrainLes '21; September 27, 2021; Virtual Event p. 173-186 URL: https://dl.acm.org/doi/abs/10.1007/978-3-031-09002-8_16 [doi: 10.1007/978-3-031-09002-8_16]
37. Jia Q, Shu H. BiTr-Unet: a CNN-transformer combined network for MRI brain tumor segmentation. In: Proceedings of the 7th International Workshop on Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. 2021 Presented at: BrainLes '21; September 27, 2021; Virtual Event p. 3-14 URL: https://link.springer.com/chapter/10.1007/978-3-031-09002-8_1 [doi: 10.1007/978-3-031-09002-8_1]
38. Hatamizadeh A, Tang Y, Nath V, Yang D, Myronenko A, Landman B, et al. UNETR: transformers for 3D medical image segmentation. In: Proceedings of the 2022 IEEE/CVF Winter Conference on Applications of Computer Vision. 2022 Presented at: WACV '22; January 4-8, 2022; Waikoloa, HI p. 1748-1758 URL: https://ieeexplore.ieee.org/document/9706678/authors [doi: 10.1109/wacv51458.2022.00181]
39. Liu J, Zheng J, Jiao G. Transition net: 2D backbone to segment 3D brain tumor. Biomed Signal Process Control 2022 May;75:103622 [FREE Full text] [doi: 10.1016/j.bspc.2022.103622]
40. Sagar A. ViTBIS: vision transformer for biomedical image segmentation. In: Proceedings of the 2021 Clinical Image-Based Procedures, Distributed and Collaborative Learning, Artificial Intelligence for Combating COVID-19 and Secure and Privacy-Preserving Machine Learning: 10th Workshop, CLIP 2021, Second Workshop, DCL 2021, First Workshop, LL-COVID19 2021, and First Workshop and Tutorial, PPML 2021, Held in Conjunction with MICCAI 2021. 2021 Presented at: MICCAI '21; September 27-October 1, 2021; Strasbourg, France p. 34-45 URL: https://dl.acm.org/doi/10.1007/978-3-030-90874-4_4 [doi: 10.1007/978-3-030-90874-4_4]
41. Dhamija T, Gupta A, Gupta S, Anjum, Katarya R, Singh G. Semantic segmentation in medical images through transfused convolution and transformer networks. Appl Intell (Dordr) 2023 Apr 25;53(1):1132-1148 [FREE Full text] [doi: 10.1007/s10489-022-03642-w] [Medline: 35498554]
42. Nalawade SS, Ganesh C, Wagner BC, Reddy D, Das Y, Yu FF, et al. Federated learning for brain tumor segmentation using mri and transformers. In: Proceedings of the 2021 Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 7th International Workshop. 2021 Presented at: BrainLes '21; September 27, 2021; Virtual Event p. 444-454 URL: https://dl.acm.org/doi/10.1007/978-3-031-09002-8_39 [doi: 10.1007/978-3-031-09002-8_39]
43. Pinaya WH, Tudosiu PD, Gray R, Rees G, Nachev P, Ourselin S, et al. Unsupervised brain imaging 3D anomaly detection and segmentation with transformers. Med Image Anal 2022 Jul;79:102475 [FREE Full text] [doi: 10.1016/j.media.2022.102475] [Medline: 35598520]
44. Cheng J, Liu J, Kuang H, Wang J. A fully automated multimodal MRI-based multi-task learning for glioma segmentation and IDH genotyping. IEEE Trans Med Imaging 2022 Jun;41(6):1520-1532 [FREE Full text] [doi: 10.1109/tmi.2022.3142321]
45. Huang J, Fang Y, Wu Y, Wu H, Gao Z, Li Y, et al. Swin transformer for fast MRI. Neurocomputing 2022 Jul;493:281-304 [FREE Full text] [doi: 10.1016/j.neucom.2022.04.051]

46. Xu X, Prasanna P. Brain cancer survival prediction on treatment-naïve MRI using deep anchor attention learning with vision transformer. In: Proceedings of the 19th International Symposium on Biomedical Imaging. 2022 Presented at: ISBI '22; March 28-31, 2022; Kolkata, India p. 1-5 URL: https://ieeexplore.ieee.org/document/9761515 [doi: 10.1109/isbi52829.2022.9761515]
47. Li H, Huang J, Li G, Liu Z, Zhong Y, Chen Y, et al. View-disentangled transformer for brain lesion detection. In: Proceedings of the 9th International Symposium on Biomedical Imaging. 2022 Presented at: ISBI '22; March 28-31, 2022; Kolkata, India p. 1-5 URL: https://ieeexplore.ieee.org/document/9761542/authors [doi: 10.1109/isbi52829.2022.9761542]
48. Zhou X, Tang C, Huang P, Tian S, Mercaldo F, Santone A. ASI-DBNet: an adaptive sparse interactive resnet-vision transformer dual-branch network for the grading of brain cancer histopathological images. Interdiscip Sci 2023 Mar 09;15(1):15-31. [doi: 10.1007/s12539-022-00532-0] [Medline: 35810266]
49. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the 2021 International Conference on Computer Vision. 2021 Presented at: ICCV '21; October 11-17, 2021; Montreal, QC p. 9992-10002 URL: https://ieeexplore.ieee.org/document/9710580/authors [doi: 10.1109/iccv48922.2021.00986]
50. RSNA-ASNR-MICCAI Brain Tumor Segmentation (BraTS) challenge 2021. University of Pennsylvania. URL: http://braintumorsegmentation.org/ [accessed 2023-11-07]
51. Multimodal brain tumor segmentation challenge 2020: data. University of Pennsylvania. URL: https://www.med.upenn.edu/cbica/brats2020/data.html [accessed 2023-11-07]
52. Multimodal brain tumor segmentation challenge 2019. University of Pennsylvania. URL: https://www.med.upenn.edu/cbica/brats-2019/ [accessed 2023-11-07]
53. Multimodal brain tumor segmentation challenge 2018. University of Pennsylvania. URL: https://www.med.upenn.edu/cbica/brats-2018/ [accessed 2023-11-07]
54. Home page. Medical Segmentation Decathlon. URL: http://medicaldecathlon.com/ [accessed 2023-11-07]
55. Brain MRI segmentation. Kaggle. URL: https://www.kaggle.com/datasets/mateuszbuda/lgg-mri-segmentation [accessed 2023-11-07]
56. Home page. UK Biobank Limited. URL: https://www.ukbiobank.ac.uk/ [accessed 2023-11-07]
57. Tools and database. Laboratory of Imaging Technologies. URL: https://lit.fe.uni-lj.si/tools.php?lang=eng [accessed 2023-11-07]
58. Multi-channel MR image: reconstruction challenge (MC-MRREC). Calgary Campinas Dataset Blog. URL: https://sites.google.com/view/calgary-campinas-dataset/mr-reconstruction-challenge [accessed 2023-11-07]
59. Isensee F, Jaeger PF, Kohl SA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods 2021 Feb;18(2):203-211. [doi: 10.1038/s41592-020-01008-z] [Medline: 33288961]
60. Chen J, Lu Y, Yu Q, Lou X, Adeli E, Wang Y, et al. TransUNet: transformers make strong encoders for medical image segmentation. arXiv. Preprint posted online February 8, 2021 2021 [FREE Full text]
61. Zheng S, Lu J, Zhao H, Zhu X, Luo Z, Wang Y, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021 Presented at: CVPR '21; June 19-25, 2021; Nashville, TN p. 6877-6886 URL: https://ieeexplore.ieee.org/document/9578646/ [doi: 10.1109/cvpr46437.2021.00681]
62. Liu Z, Ning J, Wei Y, Zhang Z, Lin S, Hu H. Video swin transformer. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022 Presented at: CVPR '22; June 19-24, 2022; New Orleans, LA p. 3192-3201 URL: https://ieeexplore.ieee.org/document/9878941 [doi: 10.1109/cvpr52688.2022.00320]
63. Wu H, Xiao B, Codella N, Liu M, Dai X, Yuan L, et al. CvT: introducing convolutions to vision transformers. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. 2021 Presented at: ICCV '21; October 11-17, 2021; Montreal, QC p. 22-31 URL: https://ieeexplore.ieee.org/document/9710031 [doi: 10.1109/iccv48922.2021.00009]
64. Karpov P, Godin G, Tetko IV. Transformer-CNN: Swiss knife for QSAR modeling and interpretation. J Cheminform 2020 Mar 18;12(1):17 [FREE Full text] [doi: 10.1186/s13321-020-00423-w] [Medline: 33431004]

Abbreviations

AI: artificial intelligence
BraTS: Brain Tumor Segmentation
CNN: convolutional neural network
GPU: graphics processing unit
IDH: isocitrate dehydrogenase
MRI: magnetic resonance imaging
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
ViT: vision transformer

Edited by A Benis; submitted 20.03.23; peer-reviewed by DK Ahmad, SQ Yoong; comments to author 11.05.23; revised version received 02.07.23; accepted 12.07.23; published 17.11.23.

Please cite as:
Ali H, Qureshi R, Shah Z
Artificial Intelligence–Based Methods for Integrating Local and Global Features for Brain Cancer Imaging: Scoping Review
JMIR Med Inform 2023;11:e47445
URL: https://medinform.jmir.org/2023/1/e47445
doi:10.2196/47445
PMID:37976086

Original Paper

Machine Learning Models for Blood Glucose Level Prediction in Patients With Diabetes Mellitus: Systematic Review and Network Meta-Analysis

Kui Liu[1*], MMS; Linyi Li[1*], MMS; Yifei Ma[1], MMS; Jun Jiang[1], MD; Zhenhua Liu[1], MD; Zichen Ye[1], MMS; Shuang Liu[1], MBA; Chen Pu[1], MMS; Changsheng Chen[2], MD; Yi Wan[1], MD

[1]Department of Health Service, Air Force Medical University, Xi'an, Shaanxi, China

[2]Department of Health Statistics, Air Force Medical University, Xi'an, Shaanxi, China

[*]these authors contributed equally

Corresponding Author:
Yi Wan, MD
Department of Health Service
Air Force Medical University
No 169, Changle West Road, Xincheng District
Xi'an, Shaanxi, 710032
China
Phone: 86 17391928966
Fax: 86 29 8471267
Email: wanyi@fmmu.edu.cn

Abstract

Background: Machine learning (ML) models provide more choices to patients with diabetes mellitus (DM) to more properly manage blood glucose (BG) levels. However, because of numerous types of ML algorithms, choosing an appropriate model is vitally important.

Objective: In a systematic review and network meta-analysis, this study aimed to comprehensively assess the performance of ML models in predicting BG levels. In addition, we assessed ML models used to detect and predict adverse BG (hypoglycemia) events by calculating pooled estimates of sensitivity and specificity.

Methods: PubMed, Embase, Web of Science, and Institute of Electrical and Electronics Engineers Explore databases were systematically searched for studies on predicting BG levels and predicting or detecting adverse BG events using ML models, from inception to November 2022. Studies that assessed the performance of different ML models in predicting or detecting BG levels or adverse BG events of patients with DM were included. Studies with no derivation or performance metrics of ML models were excluded. The Quality Assessment of Diagnostic Accuracy Studies tool was applied to assess the quality of included studies. Primary outcomes were the relative ranking of ML models for predicting BG levels in different prediction horizons (PHs) and pooled estimates of the sensitivity and specificity of ML models in detecting or predicting adverse BG events.

Results: In total, 46 eligible studies were included for meta-analysis. Regarding ML models for predicting BG levels, the means of the absolute root mean square error (RMSE) in a PH of 15, 30, 45, and 60 minutes were 18.88 (SD 19.71), 21.40 (SD 12.56), 21.27 (SD 5.17), and 30.01 (SD 7.23) mg/dL, respectively. The neural network model (NNM) showed the highest relative performance in different PHs. Furthermore, the pooled estimates of the positive likelihood ratio and the negative likelihood ratio of ML models were 8.3 (95% CI 5.7-12.0) and 0.31 (95% CI 0.22-0.44), respectively, for predicting hypoglycemia and 2.4 (95% CI 1.6-3.7) and 0.37 (95% CI 0.29-0.46), respectively, for detecting hypoglycemia.

Conclusions: Statistically significant high heterogeneity was detected in all subgroups, with different sources of heterogeneity. For predicting precise BG levels, the RMSE increases with a rise in the PH, and the NNM shows the highest relative performance among all the ML models. Meanwhile, current ML models have sufficient ability to predict adverse BG events, while their ability to detect adverse BG events needs to be enhanced.

Trial Registration: PROSPERO CRD42022375250; https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=375250

(JMIR Med Inform 2023;11:e47833) doi:10.2196/47833

KEYWORDS

machine learning; diabetes; hypoglycemia; blood glucose; blood glucose management

Introduction

Diabetes mellitus (DM) has become one of the most serious health problems worldwide [1], with more than 463 million (9.3%) patients in 2019; this number is predicted to reach 700 million (10.9%) in 2045 [2], which has resulted in growing concerns about the negative impacts on patients' lives and the increasing burden on the health care system [3]. Furthermore, previous studies have shown that without appropriate medical care, DM can lead to multiple long-term complications in blood vessels, eyes, kidneys, feet (ulcers), and nerves [4-7]. Adverse blood glucose (BG) events are one of the most common short-term complications, including hypoglycemia with BG<70 mg/dL and hyperglycemia with BG>180 mg/dL. Hyperglycemia in patients with DM may lead to lower limb occlusions and extremity nerve damage, further leading to decay, necrosis, and local or whole-foot gangrene, even requiring amputation [8,9]. Hypoglycemia can cause serious symptoms, including anxiety, palpitation, and confusion in a mild scenario and seizures, coma, and even death in a severe scenario [10,11]. Thus, there is an imminent need for preventing adverse BG events.

Machine learning (ML) models use statistical techniques to provide computers with the ability to complete assignments by training themselves without being explicitly programmed [12]. However, ML models for managing BG requires huge amounts of BG data, which cannot be satisfied by the multiple data points generated by the traditional finger-stick glucose meter [13]. With the introduction of the continuous glucose monitoring (CGM) device, which typically produces a BG reading every 5 minutes all day long, the size of the data set of BG readings is sufficient to be used in ML models [14].

Recently, there has been an immense surge in using ML technologies for predicting DM complications. Regarding BG management, previous studies have developed different types of ML models, including random forest (RF) models, support vector machines (SVMs), neural network models (NNMs), and autoregression models (ARMs), using CGM data, electronic health records (EHRs), electrocardiograph (ECG), electroencephalograph (EEG), and other information (ie, biochemical indicators, insulin intake, exercise, and meals) [10,15-20]. However, the performance of different models in these studies was not quite consistent. For instance, in terms of BG level prediction, Prendin et al [21] showed that the SVM achieved a lower root mean square error (RMSE) than the ARM, while Zhu et al [22] showed a different result.

Therefore, this meta-analysis aimed to comprehensively assess the performance of ML models in BG management in patients with DM.

Methods

Search Strategy and Study Selection

The study protocol has been registered in the international prospective register of systematic reviews (PROSPERO; registration ID: CRD42022375250). Studies on BG levels or adverse BG event prediction or detection using ML models were eligible, with no restrictions on language, investigation design, or publication status. PubMed, Embase, Web of Science, and Institute of Electrical and Electronics Engineers (IEEE) Explore databases were systematically searched from inception to November 2022. Keywords used for study repository searches were ("machine learning" OR "artificial intelligence" OR "logistic model" OR "support vector machine" OR "decision tree" OR "cluster analysis" OR "deep learning" OR "random forest") AND ("hypoglycemia" OR "hyperglycemia" OR "adverse glycemic events") AND ("prediction" OR "detection"). Details regarding the search strategies are summarized in Multimedia Appendix 1. Manual searches were added to review reference lists in relevant studies.

Selection Criteria

Inclusion criteria were as follows: (1) participants in the studies were diagnosed with DM; (2) study endpoints were hypoglycemia, hyperglycemia, or BG levels; (3) the studies established at least 2 or more types of ML models for prediction of BG levels and 1 or more types of ML models for prediction or detection of adverse BG events; (4) the studies reported the performance of ML models with statistical or clinical metrics; (5) the studies contained the development and validation of ML models; and (6) study outcomes were means (SDs) of performance metrics of test data for prediction of BG levels and sensitivity and specificity of test data for prediction or detection of adverse BG events.

Exclusion criteria were as follows: (1) studies did not report on the derivation of ML models, (2) studies were based only on physiological or control-oriented ML models, (3) studies could not reproduce true positives, true positives, false negatives, and false positives for prediction or detection of adverse BG events, (4) studies were reviews, systematic reviews, animal studies, or irretrievable and repetitive papers, and (5) studies had unavailable full text or outcome metrics.

Authors KL and LYL screened and selected studies independently based on the criteria mentioned before. Authors KL and YM extracted and recorded the data from the selected studies. Conflicts were resolved by reaching a consensus. The study strictly followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) statement (Multimedia Appendix 2) [23-25].

Data Extraction and Management

Two reviewers independently carried out data extraction and quality assessment. If a single study included more than 1 extractable test results for the same ML model, the best result was extracted. If a single study included 2 or more models, the performance metrics of each model were extracted. For studies predicting BG levels, RMSEs based on different prediction horizons (PHs) were extracted. For studies predicting or detecting adverse BG events, the sensitivity, specificity, and

precision of reproducing the 2×2 contingency table were extracted.

Specifically, the following information was extracted:

- General characteristics: first author, publication year, country, data source, and study purpose (ie, predicting or detecting hypoglycemia)
- Experimental information: participants (type of DM, type 1 or 2), sample size (patients, data points, and hypoglycemia), demographic information, models, study place and time, model parameters (ie, input and PHs), model performance metrics, threshold of BG levels for hypoglycemia, and reference (ie, finger-stick)

Methodological Quality Assessment of Included Reviews

The Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) tool was applied to assess the quality of included studies based on patient selection (5 items), index test (3 items), reference standard (4 items), and flow and timing (4 items). All 4 domains were used for assessing the risk of bias, and the first 3 domains were used to assess the consensus of applicability. Each domain has 1 query in relation to the risk of bias or applicability consisting of 7 questions [26].

Data Synthesis and Statistical Analysis

The performance metrics of ML models used to predict BG levels, predict adverse BG events, and detect adverse BG events were assessed independently. The performance metrics were the RMSE of ML models in predicting BG levels and the sensitivity and specificity of ML models in predicting or detecting adverse BG events. A network meta-analysis was conducted for BG level–based studies to assess the global and local inconsistency between studies and plotted the surface under the cumulative ranking (SUCRA) curve of every model to calculate relative ranks. For event-based studies, pooled sensitivity, specificity, the positive likelihood ratio (PLR), and the negative likelihood ratio (NLR) with 95% CIs were calculated. Study heterogeneity was assessed by calculating I^2 values based on multivariate random-effects meta-regression that considered within- and between-study correlation and classifying them into quartiles (0% to <25% for low, 25% to <50% for low-to-moderate, 50% to <75% for moderate-to-high, and >75% for high heterogeneity) [27,28]. Furthermore, meta-regression was used to evaluate the source of heterogeneity for both BG level–based and adverse event–based studies. The summary receiver operating characteristic (SROC) curve of every model was also used to evaluate the overall sensitivity and specificity. Publication bias was assessed using the Deek funnel plot asymmetry test.

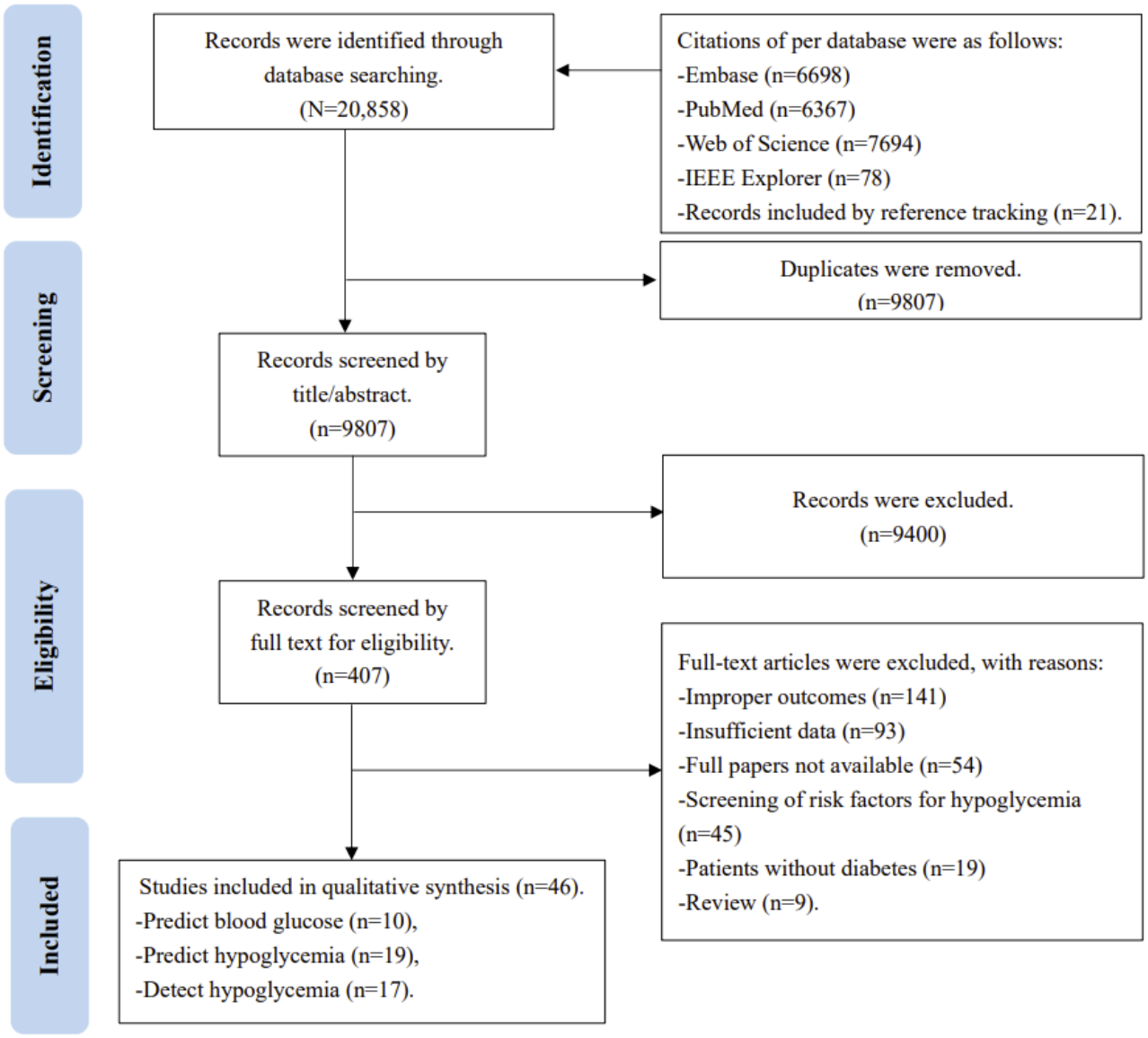
Furthermore, BG level–based studies were divided into 4 subgroups based on different PHs (15, 30, 45, 60 minutes), and adverse event–based studies were analyzed using different types of models (ie, NNM, RF, and SVM). A 2-sided *P* value of <.05 was considered statistically significant. All statistical analyses were performed using Stata 17 (Stata Corp) and Review Manager (RevMan; Cochrane) version 5.3.

Results

Search Results

A total of 20,837 studies were identified through systematically searching the predefined electronic databases; these also included 21 studies found using reference tracking [10,29-48]. Of the 20,837 studies, 9807 (47.06%) were retained after removing duplicates. After screening titles and abstracts, 9400 (95.85%) studies were excluded owing to reporting irrelevant topics or no predefined outcomes. The remaining 407 (4.15%) studies were retrieved for full-text evaluation. Of these, 361 (88.7%) studies were excluded for various reasons, and therefore 46 (11.3%) studies were included in the final meta-analysis (Figure 1).

Figure 1. Flow diagram of identifying and including studies. IEEE: Institute of Electrical and Electronics Engineers.



Description of Included Studies

As studies on hyperglycemia were insufficient for analysis, we selected studies on hypoglycemia to assess the ability of ML models to predict adverse BG events. In total, the 46 studies included 28,775 participants: n=428（1.49%）for predicting BG levels, n=28,138 (97.79%) for predicting adverse BG events, and n=209 (0.72%) for detecting adverse BG events. Of the 46 studies, 10 (21.7%) [20-22,49-55] predicted BG levels (Table 1), 19 (41.3%) [15,29-39,47,48,56-60] predicted adverse BG events (Table 2), and the remaining 17 (37%) [10,16,40-46,61-68] detected adverse BG events (Table 3).

Table 1. Baseline characteristics of BG[a] level-based studies (N=10).

First author (year), country	Data source	Sample size		Demographic information	Object; setting	Model; PH[b] (minutes); input	Performance metrics
		Patients, n	Data points, n				
Pérez-Gandía (2010), Spain [20]	CGM[c] device	15	728	—[d]	T1DM[e]; out	Models: NNM[f], ARM[g] PH: 15, 30 Input: CGM data	RMSE[h], delay
Prendin (2021) United States [21]	CGM device	Real (n=141)	350,000	Age	T1DM; out	ARM, autoregressive moving average (ARMA), autoregressive integrated moving average (ARIMA), SVM[i], RF[j] feed-forward neural network (fNN), long short-term memory (LSTM) PH: 30 Input: CGM data	RMSE, coefficient of determination (COD) sensibility, delay, precision F_1 score, time gain
Zhu (2020) England [22]	Ohio T1DM, UVA/Padova T1D	Real (n=6), simulated (n=10)	1,036,800	—	T1DM; out	DRNN[k], NNM, SVM, ARM PH:30 Input: BG level, meals, exercise, meal times	RMSE, mean absolute relative difference (MARD) time gain
D'Antoni (2020), Italy [49]	Ohio T1DM	6	—	Age, sex ratio	T1DM; out	ARJNN[l], RF, SVM, autoregression (AR), one symbolic model (SAX), recurrent neural network (RNN), one neural network model (NARX), jump neural network (JNN), delayed feed-forward neural network model (DFFNN) PH: 15, 30 Input: CGM data	RMSE
Amar (2020), Israel [50]	CGM device, insulin pump	141	1,592,506	Age, sex ratio, weight, BMI, duration of DM	T1DM; in	ARM, gradually connected neural network (GCN), fully connected (FC [neural network]), light gradient boosting machine (LCBM), RF PH: 30, 60 Input: CGM data	RMSE, Clarke error grid (CEG)
Li (2020), England [51]	UVA/Padova T1D	Simulated (n=10)	51,840	—	T1DM; out	GluNet, NNM, SVM, latent variable with exogenous input (LVX), ARM PH: 30, 60 Input: BG level, meals, exercise	RMSE, MARD, time lag
Zecchin (2012), Italy [52]	UVA/Padova T1D, CGM device	Simulated (n=20), real (n=15)	—	—	T1DM; out	Neural network–linear prediction algorithm (NN-LPA), NN, ARM PH: 30 Input: meals, insulin	RMSE, energy of second-order differences (ESOD), time gain, J index
Mohebbi (2020), Denmark [53]	Cornerstones4Care platform	Real (n=50	—	—	T1DM; in	LSTM, ARIMA PH: 15, 30, 45, 60, 90	RMSE, MAE
Daniels (2022), England [54]	CGM device	Real (n=12)	—	Sex ratio	T1DM; out	Convolutional recurrent neural network (CRNN), SVM PH: 30, 45, 60, 90, 120 Input: BG level, insulin, meals, exercise	RMSE, MAE, CEG, time gain
Alfian (2020), Korea [55]	CGM device	Real (n=12)	26,723	—	—	SVM, k-nearest neighbor k-nearest neighbor (kNN), DT[m], RF, AdaBoost, XGBoost[n], NNM PH: 15, 30 Input: CGM data	RMSE, glucose-specific root mean square error (gRMSE), R2 score, mean absolute percentage error (MAPE)

[a]BG: blood glucose.

[b]PH: prediction horizon.

[c]CGM: continuous glucose monitoring.

[d]Not applicable.

[e]T1DM: type 1 diabetes mellitus.

[f]NNM: neural network model.

[g]ARM: autoregression model.

[h]RMSE: root mean square error.
[i]SVM: support vector machine.
[j]RF: random forest.
[k]DRNN: dilated recurrent neural network.
[l]ARJNN: ARTiDe jump neural network.
[m]DT: decision tree.
[n]XGBoost: Extreme Gradient Boosting.

Table 2. Baseline characteristics of studies predicting adverse BG[a] events (N=19).

First author (year), country	Data source	Sample size			Object; setting	Model	Time	Age (years), mean (SD)/range	Threshold
		Patients, n	Data points, n	Hypoglycemia, n					
Pils (2014), United States [39]	CGM[b] device	2	2518	152	T1DM[c]; out	SVM[d]	All	—[e]	3.9
Seo (2019), Korea [15]	CGM device	104	7052	412	DM[f]; out	RF[g], SVM, k-nearest neighbor (kNN), logistic regression (LR)	Postprandial	52	3.9
Parcerisas (2022), Spain [29]	CGM device	10	67	22	T1DM; out	SVM	Nocturnal	31.8 (SD 16.8)	3.9
Stuart (2017), Greece [30]	EHRs[h]	9584	—	1327	DM; in	Multivariable logistic regression (MLR)	All	—	4
Bertachi (2020), Spain [31]	CGM device	10	124	39	T1DM; out	SVM	Nocturnal	31.8 (SD 16.8)	3.9
Elhadd (2020), Qatar [32]	—	13	3918	172	T2DM; out	XGBoost[i]	All	35-63	—
Mosquera-Lopez (2020), United States [33]	CGM device	10	117	17	T1DM; out	SVM	Nocturnal	33.7 (SD 5.8)	3.9
Mosquera-Lopez (2020), United States [33]	CGM device	20	2706	258	T1DM; out	SVM	Nocturnal	—	3.9
Ruan (2020), England [34]	EHRs	17,658	3276	703	T1DM; in	XGBoost, LR, stochastic gradient descent (SGD), kNN, DT[j], SVM, quadratic discriminant analysis (QDA), RF, extra tree (ET), linear discriminant analysis (LDA), AdaBoost, bagging	All	66 (SD 18)	4
Güemes (2020), United States [35]	CGM device	6	55	6	T1DM; out	SVM	Nocturnal	40-60	3.9
Jensen (2020), Denmark [36]	CGM device	463	921	79	T1DM; out	LDA	Nocturnal	43 (SD 15)	3
Oviedo (2019), Spain [37]	CGM device	10	1447	420	T1DM; out	SVM	Postprandial	41 (SD 10)	3.9
Toffanin (2019), Italy [38]	CGM device	20	7096	36	T1DM; out	Individual model-based	All	46	3.9

First author (year), country	Data source	Sample size			Object; setting	Model	Time	Age (years), mean (SD)/range	Threshold
		Patients, n	Data points, n	Hypoglycemia, n					
Bertachi (2018), United States [47]	CGM device	6	51	6	T1DM; out	NNM[k]	Nocturnal	40-60	3.9
Eljil (2014), United Arab Emirates [48]	CGM device	10	667	100	T1DM; out	Bagging	All	25	3.3
Dave (2021), United States [56]	CGM device	112	546,640	12,572	T1DM; out	RF	All	12.67 (SD 4.84)	3.9
Marcus (2020), Israel [57]	CGM device	11	43,533	5264	T1DM; out	Kernel ridge regression (KRR)	All	18-39	3.9
Reddy (2019), United States [58]	—	55	90	29	T1DM; out	RF	—	33 (SD 6)	3.9
Sampath (2016), Australia [59]	—	34	150	40	T1DM; out	Ranking aggregation (RA)	Nocturanl	—	—
Sudharsan (2015), United States [60]	—	—	839	428	T2DM; out	RF	All	—	3.9

[a]BG: blood glucose.

[b]CGM: continuous glucose monitoring.

[c]T1DM: type 1 diabetes mellitus.

[d]SVM: support vector machine.

[e]Not applicable.

[f]DM: diabetes mellitus.

[g]RF: random forest.

[h]EHR: electronic health record.

[i]XGBoost: Extreme Gradient Boosting.

[j]DT: decision tree.

[k]NNM: neural network model.

Table 3. Baseline characteristics of studies detecting adverse BG[a] events (N=17).

First author (year), country	Data source	Sample size			Object; setting	Model	Time	Age (years), mean (SD)/range	Threshold
		Patients, n	Data points, n	Hypoglycemia, n					
Jin (2019), United States [10]	EHRs[b]	—[c]	4104	132	T1DM[d]; in	Linear discriminant analysis (LDA)	All	—	—
Nguyen (2013), Australia [16]	EEG[e]	5	144	76	T1DM; in	Levenberg-Marquardt (LM), genetic algorithm (GA)	All	12-18	3.3
Chan (2011), Australia [40]	CGM[f] device	16	100	52	T1DM; experimental	Feed-forward neural network (fNN)	Nocturnal	14.6 (SD 1.5)	3.3
Nguyen (2010), Australia [41]	EEG	6	79	27	T1DM; experimental	Block-based neural network (BRNN)	Nocturnal	12-18	3.3
Rubega (2020), Italy [42]	EEG	34	2516	1258	T1DM; experimental	NNM[g]	All	55 (SD 3)	3.9
Chen (2019), United States [43]	EEG	—	300	11	DM[h]; in	Logistic regression (LR)	All	—	—
Jensen (2013), Denmark [44]	CGM device	10	1267	160	T1DM; experimental	SVM[i]	All	44 (SD 15)	3.9
Skladnev (2010), Australia [45]	CGM device	52	52	11	T1DM; in	fNN	Nocturnal	16.1 (SD 2.1)	3.9
Iaione (2005), Brazil [46]	EEG	8	1990	995	T1DM; experimental	NNM	Morning	35 (SD 13.5)	3.3
Nuryani (2012), Australia [61]	ECG	5	575	133	DM; in	SVM, linear multiple regression (LMR)	All	16 (SD 0.7)	3.0
San (2013), Australia [62]	ECG	15	440	39	T1DM; in	Block-based neural network (BBNN), wavelet neural network (WNN), fNN, SVM	All	14.6 (SD 1.5)	3.3
Ling (2012), Australia [63]	ECG	16	269	54	T1DM; in	Fuzzy reasoning model (FRM), fNN, multiple regression–fuzzy inference system (MR-FIS)	Nocturnal	14.6 (SD 1.5)	3.3

First author (year), country	Data source	Sample size			Object; setting	Model	Time	Age (years), mean (SD)/range	Threshold
		Patients, n	Data points, n	Hypoglycemia, n					
Ling (2016), Australia [64]	ECG	16	269	54	T1DM; in	Extreme learning machine–based neural network (ELM-NN), particle swarm optimization–based neural network (PSO-NN), MR-FIS, LMR, fuzzy inference system (FIS)	Nocturnal	14.6 (SD 1.5)	3.3
Nguyen (2012), Australia [65]	EEG	5	44	20	T1DM; in	NNM	—	12-18	3.3
Ngo (2020), Australia [66]	EEG	8	135	53	T1DM; in	BRNN	Nocturnal	12-18	3.9
Ngo (2018), Australia [67]	EEG	8	54	26	T1DM; in	BRNN	Nocturnal	12-18	3.9
Nuryani (2010), Australia [68]	ECG	5	27	8	T1DM; experimental	Fuzzy support vector machine (FSVM), SVM	Nocturnal	16 (SD 0.7)	3.3

[a]BG: blood glucose.

[b]EHR: electronic health record.

[c]Not applicable.

[d]T1DM: type 1 diabetes mellitus.

[e]EEG: electroencephalograph.

[f]CGM: continuous glucose monitoring.

[g]NNM: neural network model.

[h]DM: diabetes mellitus.

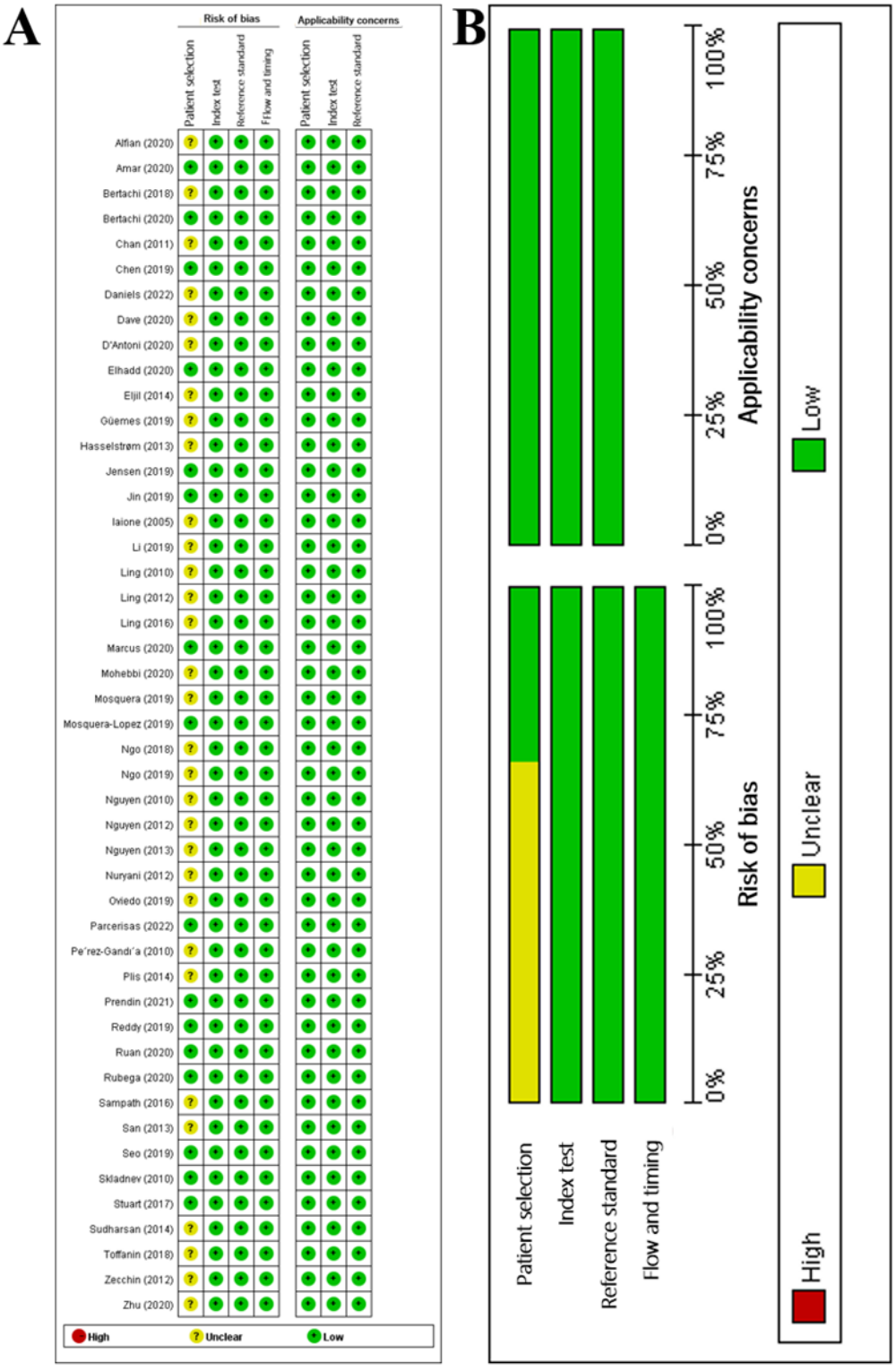
[i]SVM: support vector machine.

As shown in Tables 1-3, 40 (87%) studies [10,16,20-22,29,31,33-42,44-59,62-68] included participants with type 1 diabetes mellitus (T1DM), 2 (4.3%) studies [32,60] included participants with type 2 diabetes mellitus (T2DM), and the remaining 4 (8.7%) studies [15,30,43,61] did not specify the type of DM. Regarding the data source of ML models, CGM devices were involved in 22 (47.8%) studies [15,20,21,29,31,33,35-40,44,45,47,48,50,52,54-57], EEG signals were used in 8 (17.4%) studies [16,41-43,46,65-67], ECG signals were involved in 5 (10.9%) studies [61-64,68], EHRs were used in 3 (6.5%) studies [10,30,34], data generated by the UVA/Padova T1D simulator were used in 3 (6.5%) studies [22,51,52], the Ohio T1DM data set was used in 2 (4.3%) studies [22,49], and 4 (8.7%) studies [32,58-60] did not report the source of data. Regarding the setting of data collection, 24 (52.2%) studies [15,20-22,29,31-33,35-39,47-49,51,52,54,56-60] were conducted in an out-of-hospital setting, 13 (28.3%) studies [10,16,34,43,50,53,61-67] were conducted in an in-hospital setting, 6 (13%) studies [40-42,44,46,68] were conducted in an experimental setting, and the remaining 1 (2.2%) study [55] did not specify the environment. Regarding when adverse BG events occurred in the 36 (78.3%) adverse event–based studies, 15 (41.7%) [29,31,33,35,36,40,41,45,47,59,63,64,66-68] reported nocturnal hypoglycemia, 16 (44.4%) [10,16,30,32,34,38,39,42-44,48,56,57,60-62] were not specific about the time of day, 2 (5.6%) [15,37] reported postprandial hypoglycemia, 1 (2.8%) [46] reported morning hypoglycemia, and the remaining 2 (5.6%) [58,65] did not report the time setting. To carry out the network meta-analysis of BG level–based studies, we chose the RMSE as the outcome to be compared.

Quality Assessment of Included Studies

The quality assessment results using the QUADAS-2 tool showed that more than half of all included studies did not report the patient selection criteria in detail, which led to low-quality patient selection (Figure 2). Furthermore, the diagnosis of hypoglycemia using blood or the CGM device was considered high quality in the reference test in our study.

Figure 2. Quality assessment of included studies. Risk of bias and applicability concerns graph (A) and risk of bias and applicability concerns summary (B).



Statistical Analysis

Machine Learning Models for Predicting Blood Glucose Levels

Network meta-analysis was conducted to evaluate the performance of different ML models. For PH=30 minutes, 10 (21.7%) studies [20-22,49-55] with 32 different ML models were included, and the network map is shown in Figure 3A. The mean RMSE was 21.40 (SD 12.56) mg/dL. Statistically significant inconsistency was detected using the inconsistency test(^{2}=87.11, *P*<.001), as shown in the forest plot in Multimedia Appendix 1. Meta-regression indicated that I^2 for the RMSE was 60.75%, and the source of heterogeneity analysis showed that place and validation type were statistically significant (*P*<.001). The maximum SUCRA value was 99.1 for the dilated recurrent neural network (DRNN) model with a mean RMSE

of 7.80 (SD 0.60) mg/dL [22], whereas the minimum SUCRA value was 0.4 for 1 symbolic model with a mean RMSE of 71.4 (SD 21.9) mg/dL [49]. The relative ranks of the ML models are shown in Table 4, and the SUCRA curves are shown in Figure 4A. Publication bias was tested using the Egger test (P=.503), indicating no significant publication bias.

For PH=60 minutes, 4 (8.7%) studies [50,51,55] with 17 different ML models were included, and the network map is shown in Figure 3B. The mean RMSE was 30.01 (SD 7.23) mg/dL. Statistically significant inconsistency was detected using the inconsistency test (^{2}=8.82, P=.012), as shown in the forest plot in Multimedia Appendix 3. Meta-regression indicated that none of the sample size, reference, place, validation type, and model type was a source of heterogeneity. The maximum SUCRA value was 97.8 for the GluNet model with a mean RMSE of 19.90 (SD 3.17) mg/dL [51], while the minimum SUCRA value was 4.5 for the decision tree (DT) model with a mean RMSE of 32.86 (SD 8.81) mg/dL [55]. The relative ranks of the ML models are shown in Table 5, and the SUCRA curves are shown in Figure 4B. No significant publication bias was detected using the Egger test (P=.626).

For PH=15 minutes, 3 (6.5%) studies [20,49,55] with 14 different ML models were included, and the network map is shown in Figure 3C. The mean RMSE was 18.88 (SD 19.71) mg/dL. Statistically significant inconsistency was detected using the inconsistency test (^{2}=28.29, P<.001), as shown in the forest plot in Multimedia Appendix 4. Meta-regression showed that I^2 was 41.28%, and the model type and sample size both were the source of heterogeneity, with P=.002 and .037, respectively. The maximum SUCRA value was 99.1 for the ARTiDe jump neural network (ARJNN) model with a mean RMSE of 9.50 (SD 1.90) mg/dL [49], while the minimum SUCRA value was 0.3 for the SVM with a mean RMSE of 13.13 (SD 17.30) mg/dL [55]. The relative ranks of the ML models are shown in Table 6, and SUCRA curves are shown in Figure 4C. Statistically significant publication bias was detected using the Egger test (P=.003).

For PH=45 minutes, only 2 (4.3%) studies [54,55] with 11 different ML models were included, and the network map is shown in Figure 3D. The mean RMSE was 21.27 (SD 5.17) mg/dL. Statistically significant inconsistency was detected using the inconsistency test (^{2}=6.92, P=.009), as shown in the forest plot in Multimedia Appendix 5. Meta-regression indicated significant heterogeneity from the model type (P=.006). The maximum SUCRA value was 99.4 for the NNM with a mean RMSE of 10.65 (SD 3.87) mg/dL [55], while the minimum SUCRA value was 26.3 for the DT model with a mean RMSE of 23.35 (6.36) mg/dL [55]. The relative ranks of the ML models are shown in Table 7, and SUCRA curves are shown in Figure 4D. Statistically significant publication bias was detected using the Egger test (P<.001).

Figure 3. Network map of ML models for predicting BG levels in different PHs. PH=30 (A), 60 (B), 15 (C), and 45 minutes (D). ARIMA: autoregressive integrated moving average; ARM: autoregression model; ARMA: autoregressive moving average; ARJNN: ARTiDe jump neural network; BG: blood glucose; CRNN-MTL: convolutional recurrent neural network multitask learning; CRNN-MTL-GV: convolutional recurrent neural network multitask learning glycemic variability; CRNN-STL: convolutional recurrent neural network single-task learning; CRNN-TL: convolutional recurrent neural network transfer learning; DFFNN: delayed feed-forward neural network; DRNN: dilated recurrent neural network; DT: decision tree; FC: fully connected (neural network); fNN: feed-forward neural network; GCN: gradually connected neural network; JNN: jump neural network; kNN: k-nearest neighbor; LGBM: light gradient boosting machine; LSTM: long short-term memory; LVX: latent variable with exogenous input; ML: machine learning; NARX: one neural network model; NN-LPA: neural network–linear prediction algorithm; NNM: neural network model; PH: prediction horizon; RF: random forest; RNN: recurrent neural network; SAX: one symbolic model; SVR: support vector regression.

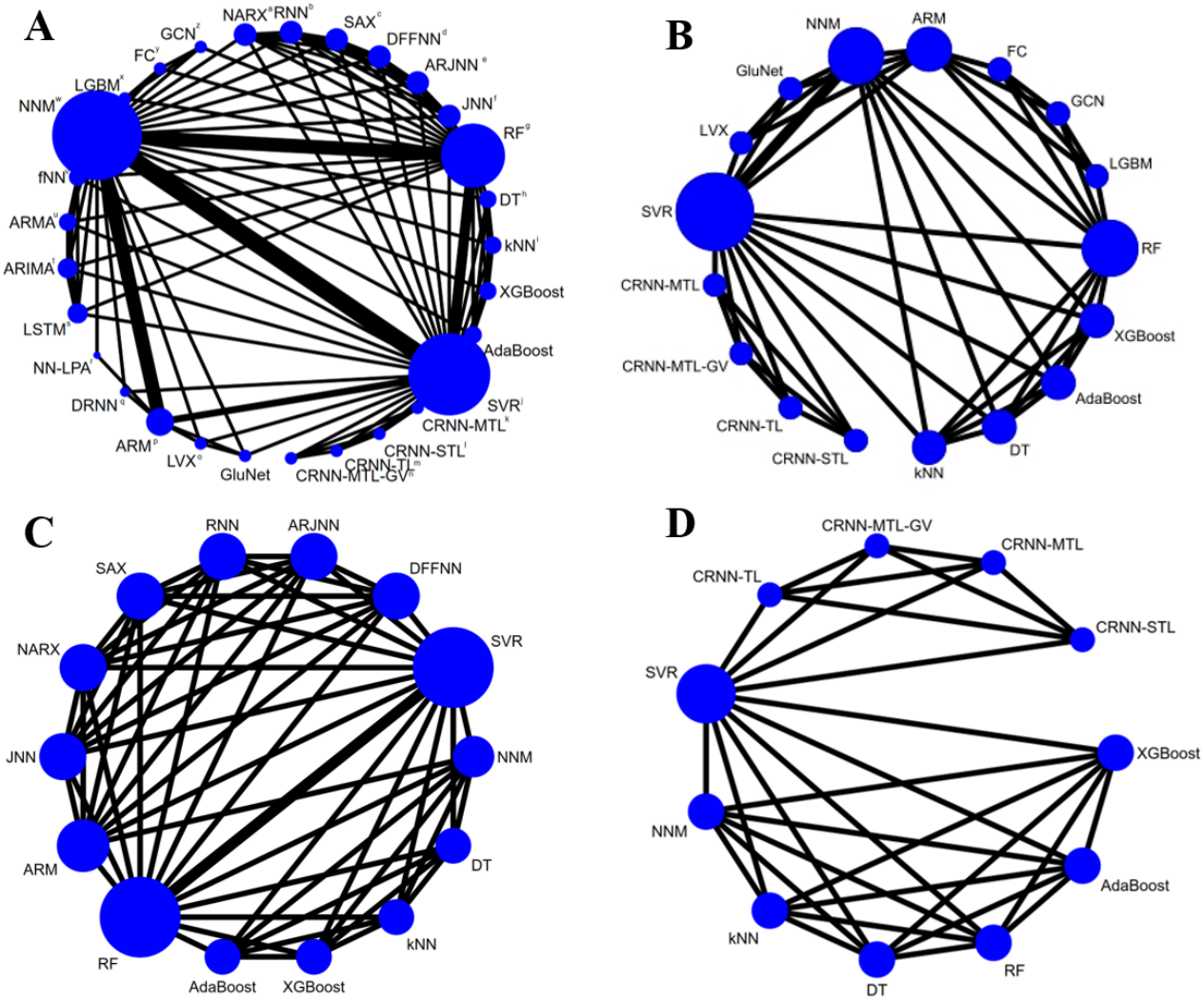


Table 4. Relative ranks of ML[a] models for predicting BG[b] levels in PH[c]=30 minutes.

ML model	SUCRA[d]	Relative rank
NNM[e]	52.0	14.4
ARM[f]	39.6	17.9
ARJNN[g]	79.5	6.8
RF[h]	6.9	27.1
SVM[i]	73.3	8.5
One symbolic model (SAX)	0.4	28.9
Recurrent neural network (RNN)	19.0	23.7
One neural network model (NARX)	3.9	27.9
Jump neural network (JNN)	36.0	18.9
Delayed feed-forward neural network model (DFFNN)	15.8	24.6
Gradually connected neural network (GCN)	41.1	17.5
Fully connected (FC [neural network])	58.1	12.7
Light gradient boosting machine (LGBM)	69.3	9.6
DRNN[j]	99.1	1.2
Autoregressive moving average (ARMA)	54.3	13.8
Autoregressive integrated moving average (ARIMA)	46.6	16.0
Feed-forward neural network (fNN)	86.3	4.8
Long short-term memory (LSTM)	69.1	9.7
GluNet	96.4	2.0
Latent variable with exogenous input (LVX)	75.2	7.9
Neural network–linear prediction algorithm (NN-LPA)	60.0	12.2
Convolutional recurrent neural network multitask learning (CRNN-MTL)	77.5	7.3
Convolutional recurrent neural network multitask learning glycemic variability (CRNN-MTL-GV)	77.2	7.4
Convolutional recurrent neural network transfer learning (CRNN-TL)	71.8	8.9
Convolutional recurrent neural network single-task learning (CRNN-STL)	52.0	14.4
k-Nearest neighbor (kNN)	26.0	21.7
DT[k]	16.2	24.5
AdaBoost	18.0	24.0
XGBoost[l]	29.2	20.8

[a]ML: machine learning.
[b]BG: blood glucose.
[c]PH: prediction horizon.
[d]SUCRA: surface under the cumulative ranking.
[e]NNM: neural network model.
[f]ARM: autoregression model.
[g]ARJNN: ARTiDe jump neural network.
[h]RF: random forest.
[i]SVM: support vector machine.
[j]DRNN: dilated recurrent neural network.
[k]DT: decision tree.
[l]XGBoost: Extreme Gradient Boosting.

Figure 4. SUCRA curves of ML models for predicting BG levels in different PHs. PH=30 (A), 60 (B), 15 (C), and 45 minutes (D). ARIMA: autoregressive integrated moving-average; ARM: autoregression model; ARMA: autoregressive moving average; ARJNN: ARTiDe jump neural network; BG: blood glucose; CRNN-MTL: convolutional recurrent neural networks multitask learning; CRNN-MTL-GV: convolutional recurrent neural networks multitask learning glycemic variability; CRNN-STL: convolutional recurrent neural networks single-task learning; CRNN-TL: convolutional recurrent neural networks transfer learning; DFFNN: delayed feed-forward neural network; DRNN: dilated recurrent neural network; DT: decision tree; FC: fully connected (neural network); fNN: feed-forward neural network; GCN: gradually connected neural network; JNN: jump neural network; kNN: k-nearest neighbor; LGBM: light gradient boosting machine; LSTM: long short-term memory; LVX: latent variable with exogenous input; ML: machine learning; NARX: one neural network model; NN-LPA: neural network–linear prediction algorithm; NNM: neural network model; PH: prediction horizon; RF: random forest; RNN: recurrent neural network; SAX: one symbolic model; SVR: support vector regression.

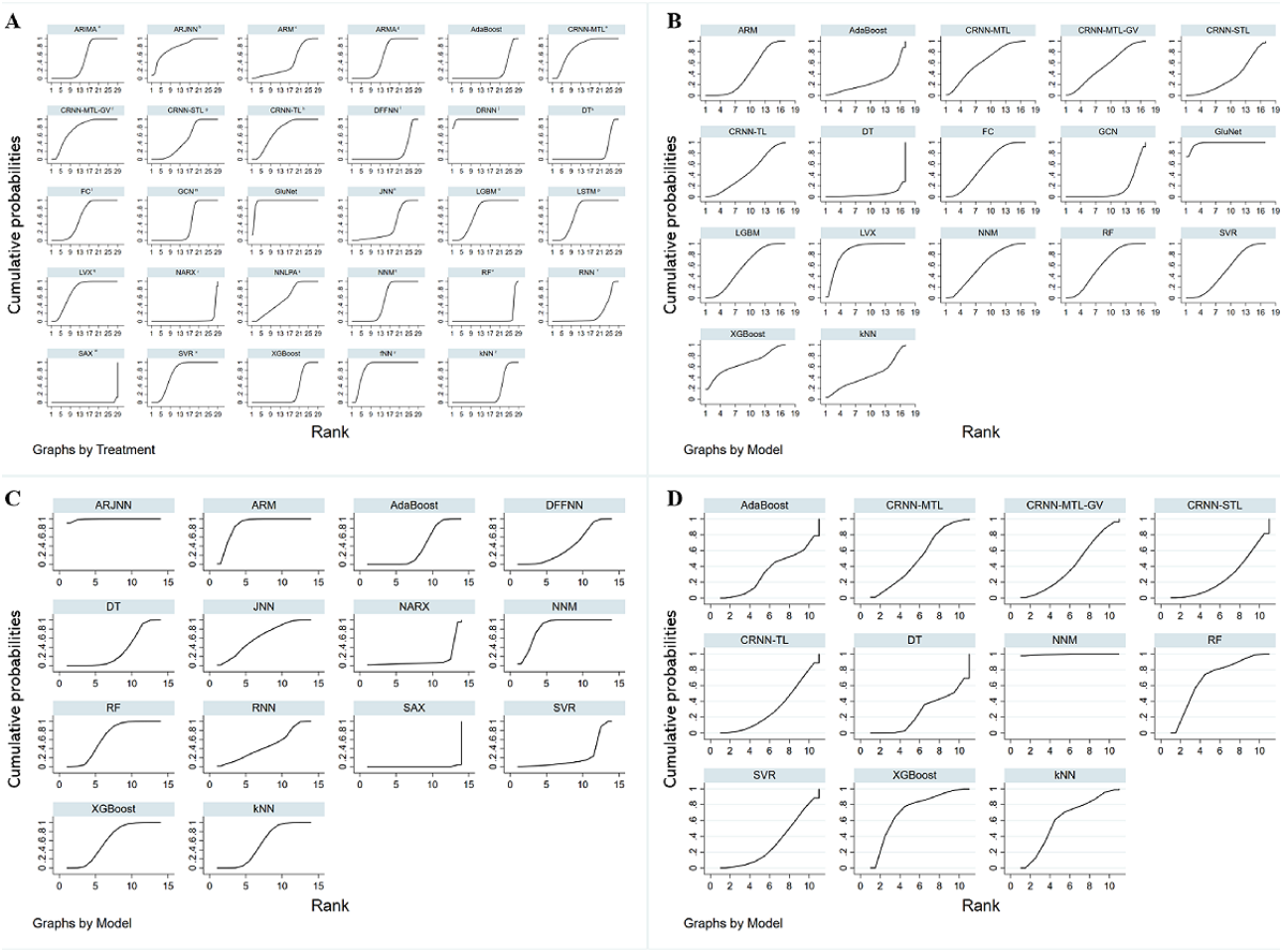


Table 5. Relative ranks of ML[a] models for predicting BG[b] levels in PH[c]=60 minutes.

ML model	SUCRA[d]	Relative rank
ARM[e]	41.0	10.4
Gradually connected neural network (GCN)	14.2	14.7
Fully connected (FC [neural network])	55.7	8.1
Light gradient boosting machine (LGBM)	56.0	8.0
RF[f]	59.7	7.5
GluNet	97.8	1.4
NNM[g]	59.9	7.4
SVM[h]	49.5	9.1
Latent variable with exogenous input (LVX)	85.9	3.3
Convolutional recurrent neural network multitask learning (CRNN-MTL)	61.4	7.2
Convolutional recurrent neural network multitask learning glycemic variability (CRNN-MTL-GV)	54.2	8.3
Convolutional recurrent neural network transfer learning (CRNN-TL)	44.5	9.9
Convolutional recurrent neural network single-task learning (CRNN-STL)	32.5	11.8
k-Nearest neighbor (kNN)	42.5	10.2
DT[i]	4.5	16.3
AdaBoost	24.1	13.1
XGBoost[j]	66.5	6.4

[a]ML: machine learning.
[b]BG: blood glucose.
[c]PH: prediction horizon.
[d]SUCRA: surface under the cumulative ranking.
[e]ARM: autoregression model.
[f]RF: random forest.
[g]NNM: neural network model.
[h]SVM: support vector machine.
[i]DT: decision tree.
[j]XGBoost: Extreme Gradient Boosting.

Table 6. Relative ranks of ML[a] models for predicting BG[b] levels in PH[c]=15 minutes.

ML model	SUCRA[d]	Relative rank
NNM[e]	84.4	3.0
ARM[f]	86.8	2.7
ARJNN[g]	99.1	1.1
RF[h]	64.6	5.6
SVM[i]	20.9	11.3
One symbolic model (SAX)	0.3	14.0
Recurrent neural network (RNN)	45.9	8.0
One neural network model (NARX)	11.8	12.5
Jump neural network (JNN)	62.2	5.9
Delayed feed-forward neural network model (DFFNN)	39.6	8.9
k-Nearest neighbor (kNN)	53.7	7.0
DT[j]	33.3	9.7
AdaBoost	36.8	9.2
XGBoost[k]	60.8	6.1

[a]ML: machine learning.
[b]BG: blood glucose.
[c]PH: prediction horizon.
[d]SUCRA: surface under the cumulative ranking.
[e]NNM: neural network model.
[f]ARM: autoregression model.
[g]ARJNN: ARTiDe jump neural network.
[h]RF: random forest.
[i]SVM: support vector machine.
[j]DT: decision tree.
[k]XGBoost: Extreme Gradient Boosting.

Table 7. Relative ranks of ML[a] models for predicting BG[b] levels in PH[c]=45 minutes.

ML model	SUCRA[d]	Relative rank
Convolutional recurrent neural network multitask learning (CRNN-MTL)	52.1	5.8
Convolutional recurrent neural network multitask learning glycemic variability (CRNN-MTL-GV)	41.8	6.8
Convolutional recurrent neural network transfer learning (CRNN-TL)	31.6	7.8
Convolutional recurrent neural network single-task learning (CRNN-STL)	27.5	8.2
SVM[e]	32.0	7.8
k-Nearest neighbor (kNN)	61.4	4.9
DT[f]	26.3	8.4
RF[g]	70.3	4.0
AdaBoost	34.1	7.6
XGBoost[h]	73.5	3.7
NNM[i]	99.4	1.1

[a]ML: machine learning.
[b]BG: blood glucose.
[c]PH: prediction horizon.
[d]SUCRA: surface under the cumulative ranking.
[e]SVM: support vector machine.
[f]DT: decision tree.
[g]RF: random forest.
[h]XGBoost: Extreme Gradient Boosting.
[i]NNM: neural network model.

Machine Learning Models for Predicting Hypoglycemia

ML models for predicting hypoglycemia (adverse BG events) involved 19 (41.3%) studies [15,29-39,47,48,56-60], with pooled estimates of 0.71 (95% CI 0.61-0.80) for sensitivity, 0.91 (95% CI 0.87-0.94) for specificity, 8.3 (95% CI 5.7-12.0) for the PLR, and 0.31 (95% CI 0.22-0.44) for the NLR. The heterogeneity between different ML models in these studies is shown in the forest plot in Figure 5, which was high for both sensitivity (I^2=100%, 95% CI 100%-100%) and specificity (I^2=100%, 95% CI 100%-100%). The SROC curve is shown in Figure 6A, with an area under the curve (AUC) of 0.91 (95% CI 0.88-0.93). According to the meta-regression results, the type of DM and time were statistically significant sources of heterogeneity for sensitivity while the type of DM, reference, data source, setting, and threshold were statistically significant sources of heterogeneity for specificity (Multimedia Appendix 6). No statistically significant publication bias was detected (*P*=.09). In addition to integral analysis for the hypoglycemia prediction model, we also carried out analysis of 4 subgroups based on the characteristics of the included studies, including the NNM, the RF, the SVM, and ensemble learning (RF, Extreme Gradient Boosting [XGBoost], bagging).

For the NNM, 3 (6.5%) studies [15,34,47] were included, with pooled estimates of 0.50 (95% CI 0.16-0.84) for sensitivity, 0.91 (95% CI 0.84-0.96) for specificity, 5.9 (95% CI 3.2-10.8) for the PLR, and 0.54 (95% CI 0.24-1.21) for the NLR. As shown in the forest plot in Figure 7A, I^2 values were 99.59% (95% CI 99.46%-99.71%) and 97.82% (95% CI 96.68%-98.86%) for sensitivity and specificity, respectively. The SROC curve is shown in Figure 6B, with an AUC of 0.90 (95% CI 0.87-0.92). Meta-regression results revealed that statistically significant heterogeneity was detected in all the factors between these studies (type of DM, reference, time, data source, setting, threshold) for sensitivity and 4 factors (reference, data source, setting, threshold) for specificity (Multimedia Appendix 7). No statistically significant publication bias was detected (*P*=.86).

For the RF, 5 (10.9%) studies [15,34,56,58,60] were included, with pooled estimates of 0.87 (95% CI 0.79-0.93) for sensitivity, 0.94 (95% CI 0.91-0.96) for specificity, 13.9 (95% CI 10.1-18.9) for the PLR, and 0.14 (95% CI 0.08-0.22) for the NLR. The forest plot in Figure 7B shows that statistically significant heterogeneity was detected in both sensitivity (I^2=98.32%, 95% CI 97.61%-99.02%) and specificity (I^2=99.41%, 95% CI 99.24%-99.58%). The SROC curve is shown in Figure 6C, with an AUC of 0.97 (95% CI 0.95-0.98). Meta-regression failed to run due to data instability or asymmetry. No statistically significant publication bias was detected (*P*=.21).

Figure 5. Sensitivity and specificity forest plots of ML models for predicting adverse BG events. The horizontal lines indicate 95% CIs. The square markers represent the effect value of a single study, and the diamond marker represents the combined results of all studies. The vertical line shows the line of no effects. BG: blood glucose; ML: machine learning.

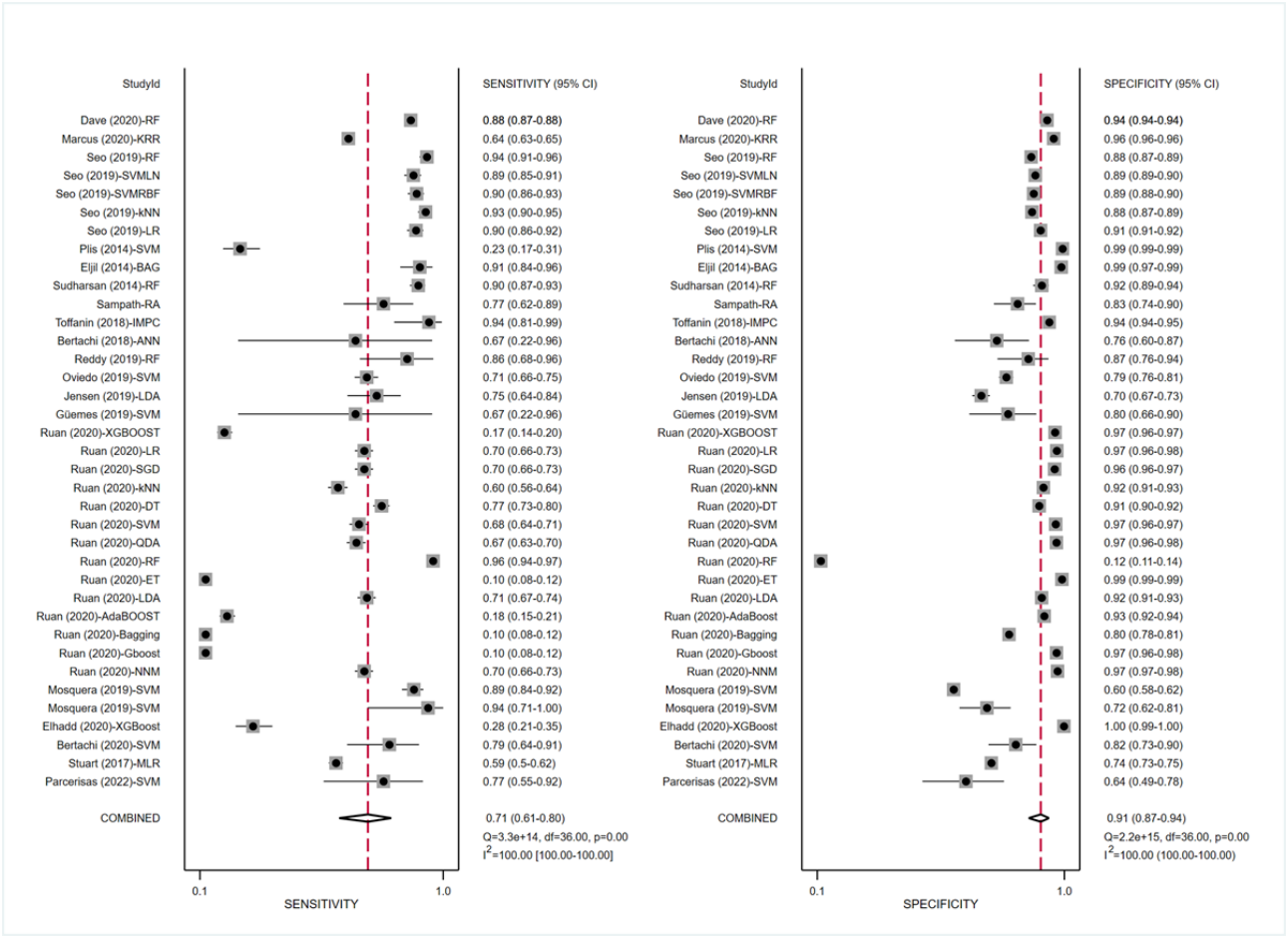


Figure 6. SROC curves of all ML algorithms (A), NNM algorithms (B), RF algorithms (C), SVM algorithms (D), and ensemble learning algorithms (E) for predicting adverse BG events. The hollow circles represent results of all studies, and the red diamonds represent the summary result of all studies. AUC: area under the curve; BG: blood glucose; ML: machine learning; NNM: neural network model; RF: random forest; SROC: summary receiver operating characteristic; SVM: support vector machine.

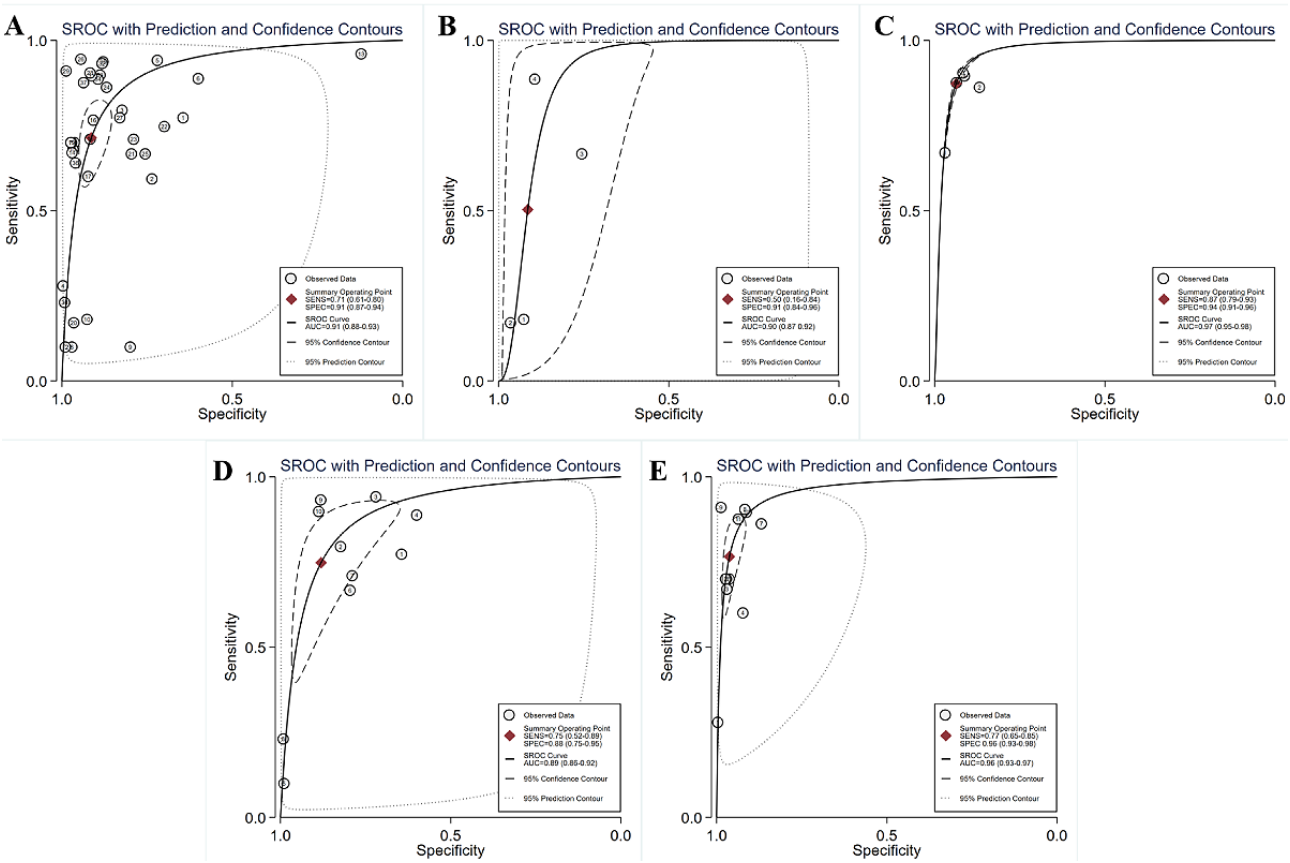
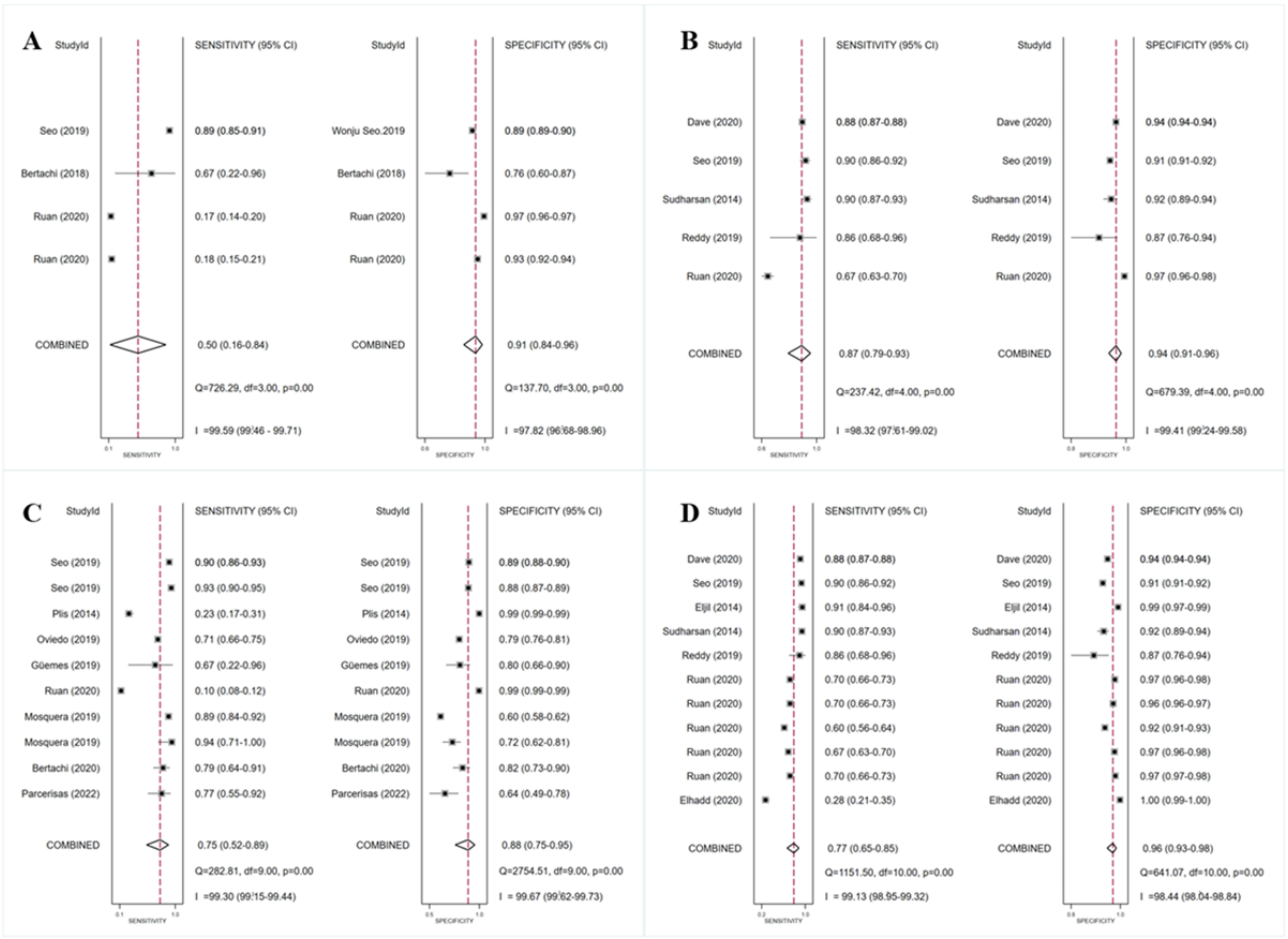


Figure 7. Sensitivity and specificity forest plots of NNM algorithms (A), RF models (B), SVM algorithms (C), and ensemble learning algorithms (D) for predicting adverse BG events. The horizontal lines indicate 95% CIs. The square markers represent the effect value of a single study, and the diamond marker represents the combined results of all studies. The vertical line shows the line of no effects. BG: blood glucose; NNM: neural network model; RF: random forest; SROC: summary receiver operating characteristic; SVM: support vector machine.



For the SVM, 8 (17.4%) studies [15,29,33-35,37,39,47] were involved, with pooled estimates of 0.75 (95% CI 0.52-0.89) for sensitivity, 0.88 (95% CI 0.75-0.95) for specificity, 6.3 (95% CI 3.4-11.7) for the PLR, and 0.29 (95% CI 0.15-0.55) for the NLR. Statistically significant heterogeneity was detected for both sensitivity (I^2=99.30%, 95% CI 99.15%-99.44%) and specificity (I^2=99.67%, 95% CI 99.62%-99.73%), as shown in Figure 7C. The SROC curve is shown in Figure 6D, with an AUC of 0.89 (95% CI 0.86-0.92). Meta-regression results showed that reference, time, data source, setting, and threshold were sources of heterogeneity for sensitivity, while reference, data source, setting, and threshold were sources of heterogeneity for specificity (Multimedia Appendix 8). Publication bias was not statistically significant (*P*=.83).

For ensemble learning models (RF, XGBoost, bagging), 7 (15.2%) studies [15,32,34,48,56,58,60] were involved, with pooled estimates of 0.77 (95% CI 0.65-0.85) for sensitivity, 0.96 (95% CI 0.93-0.98) for specificity, 20.4 (95% CI 12.5-33.3) for the PLR, and 0.24 (95% CI 0.16-0.37) for the NLR. Statistically significant heterogeneity was detected for both sensitivity (I^2=99.13%, 95% CI 98.95%-99.32%) and specificity (I^2=98.44%, 95% CI 98.04%-98.84%), as shown in Figure 7D. The SROC curve is shown in Figure 6E, with an AUC of 0.96 (95% CI 0.93-0.97). Meta-regression results showed that there was no source of heterogeneity for sensitivity, while the type of DM, setting, and threshold were sources of heterogeneity for specificity (Multimedia Appendix 9). No statistically significant publication bias was detected (*P*=.50).

Machine Learning Models for Detecting Hypoglycemia

ML models for detecting hypoglycemia (adverse BG events) involved 17 (37%) studies [10,16,40-46,61-68], with pooled estimates of 0.74 (95% CI 0.70-0.78) for sensitivity, 0.70 (95% CI 0.56-0.81) for specificity, 2.4 (95% CI 1.6-3.7) for the PLR, and 0.37 (95% CI 0.29-0.46) for the NLR. The heterogeneity between different models in these studies is shown in the forest plots in Figure 8 and was high for both sensitivity (I^2=92.80%, 95% CI 91.10%-94.49%) and specificity (I^2=99.04%, 95% CI 98.82%-99.16%). The SROC curve is shown in Figure 9A, with an AUC of 0.77 (95% CI 0.73-0.81). Based on the meta-regression results, reference, time, data source, setting, and threshold were statistically significant sources of heterogeneity for sensitivity, while reference, data source, and threshold were statistically significant sources of heterogeneity for specificity (Multimedia Appendix 9). Statistically significant publication bias was detected (*P*<.001). In addition to integral analysis for the hypoglycemia detection model, we also carried out analysis of 2 subgroups based on the characteristics of the included studies, including the NNM and the SVM.

For the NNM, 11 (23.9%) studies [40-42,45,46,62-67] were involved, with pooled estimates of 0.76 (95% CI 0.70-0.80) for sensitivity, 0.67 (95% CI 0.49-0.82) for specificity, 2.3 (95% CI 1.4-3.9) for the PLR, and 0.36 (95% CI 0.27-0.48) for the NLR. The heterogeneity between different studies is shown in the forest plot in Figure 10A and was high for both sensitivity (I^2=97.30%, 95% CI 96.62%-97.99%) and specificity (I^2=98.23%, 95% CI 97.83%-98.62%). The SROC curve is shown in Figure 9B, with an AUC of 0.78 (95% CI 0.74-0.81). Based on the of meta-regression results, reference, time, data source, setting, and threshold were statistically significant sources of heterogeneity for sensitivity, while reference and setting were statistically significant sources of heterogeneity for specificity (Multimedia Appendix 10). Statistically significant publication bias was detected (P<.001).

For the SVM, 4 (8.7%) studies [10,44,61,62] were included, with pooled estimates of 0.80 (95% CI 0.73-0.86) for sensitivity, 0.65 (95% CI 0.41-0.83) for specificity, 2.3 (95% CI 1.2-4.4) for the PLR, and 0.31 (95% CI 0.18-0.51) for the NLR. The heterogeneity between different studies is shown in the forest plot in Figure 10B and was high for both sensitivity (I^2=55.86%, 95% CI 11.96%-99.76%) and specificity (I^2=99.02%, 95% CI 98.68%-99.36%). The SROC curve is shown in Figure 9C, with an AUC of 0.81 (95% CI 0.78-0.85). Meta-regression results indicated that reference, time, data source, setting, and threshold were statistically significant sources of heterogeneity for sensitivity, while reference, data source, setting, and threshold statistically significant sources of heterogeneity for specificity (Multimedia Appendix 11). No statistically significant publication bias was detected (P=.31).

Figure 8. Sensitivity and specificity forest plots of ML models for detecting adverse BG events. The horizontal lines indicate 95% CIs. The square markers represent the effect value of a single study, and the diamond marker represents the combined results of all studies. The vertical line shows the line of no effects. BG: blood glucose; ML: machine learning.

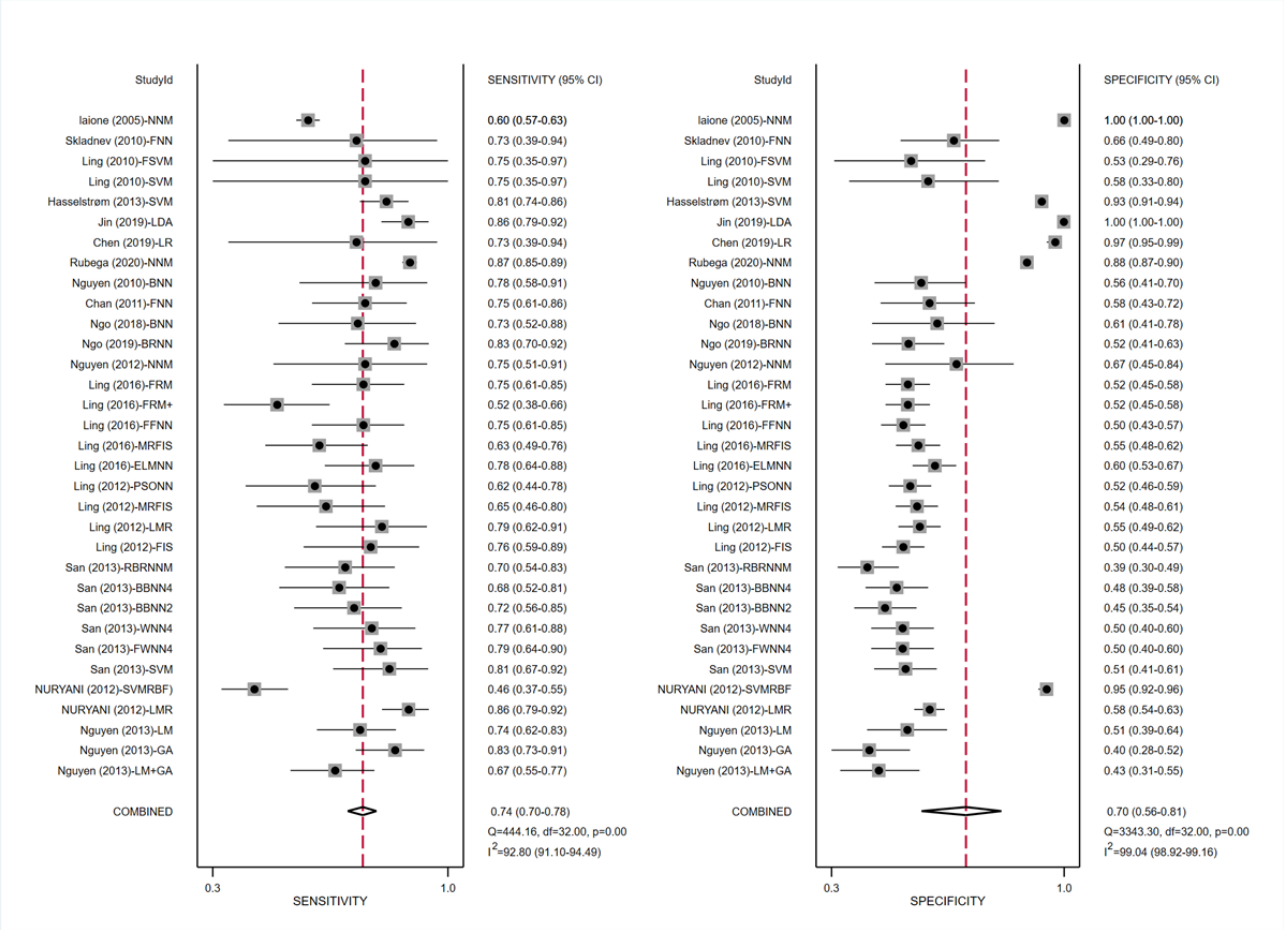


Figure 9. SROC curves of all ML algorithms (A), NNM algorithms (B), and SVM algorithms (C) for detecting adverse BG events. The hollow circles represent results of all studies, and the red diamonds represent the summary result of all studies. AUC: area under the curve; BG: blood glucose; ML: machine learning; NNM: neural network model; SROC: summary receiver operating characteristic; SVM: support vector machine.

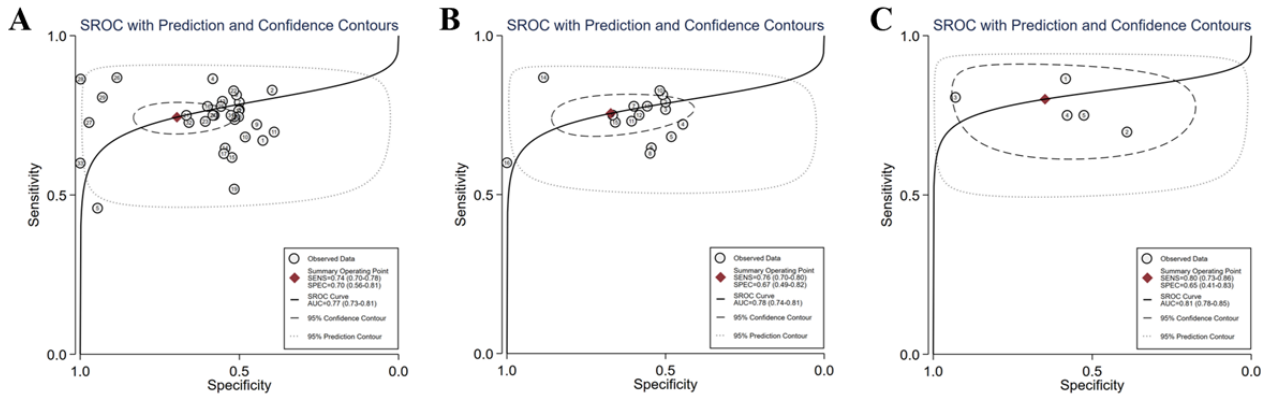
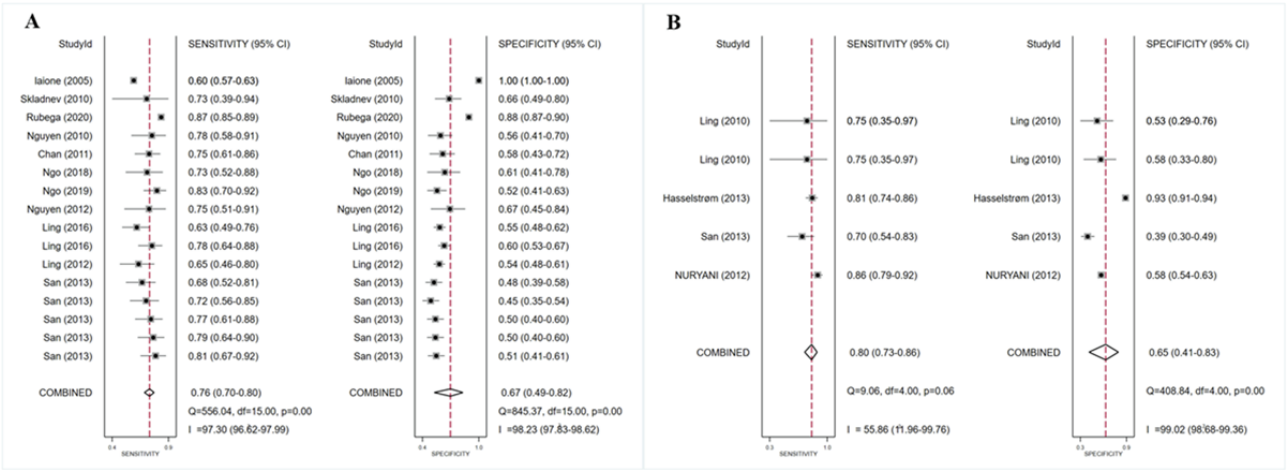


Figure 10. Sensitivity and specificity forest plots of NNM algorithms (A) and SVM algorithms (B) for detecting adverse BG events. The horizontal lines indicate 95% CIs. The square markers represent the effect value of a single study, and the diamond marker represents the combined results of all studies. The vertical line shows the line of no effects. BG: blood glucose; NNM: neural network model; SVM: support vector machine.



Discussion

Principal Findings

This meta-analysis systematically assessed the performance of different ML models in enhancing BG management in patients with DM based on 46 eligible studies. Comprehensive evidence obtained via exhaustive searching allowed us to assess the overall ability of the ML models in different scenarios, including predicting BG levels, predicting adverse BG events, and detecting adverse BG events.

Comparison to Prior Work

Obviously, the RMSE of ML models for predicting BG levels increased as the PH increased from 15 to 60 minutes, which indicates that the longer the PH, the larger the prediction error. Based on the results of relative ranking, among all the ML models for predicting BG levels, neural network–based models, including the DRNN, GluNet, ARJNN, and NNM, achieved the minimum RMSE and the maximum SUCRA in different PHs, indicting the highest relative performance. In contrast, the DT achieved the maximum RMSE and the minimum SUCRA in a PH of 60 and 45 minutes, indicating that lowest relative performance. Thus, for predicting BG levels, neural network–based algorithms might be an appropriate choice. We found that time domain features combined with historical BG levels as input can further improve the performance of NNM algorithms [49,55]. However, the quality of training data for NNMs needs to be high; therefore, the requirements during data collection and preprocessing of raw data are high [22,51].

Regarding ML models for predicting adverse BG events, the pooled sensitivity, specificity, PLR, and NLR were 0.71 (95% CI 0.61-0.80), 0.91 (95% CI 0.87-0.94), 8.3 (95% CI 5.7-12.0), and 0.31 (95% CI 0.22-0.44), respectively. According to the *Users' Guide to Medical Literature*, with regard to diagnostic tests [69], a PLR of 5-10 should be able to moderately increase the probability of persons having or developing a disease and an NLR of 0.1-0.2 should be able to moderately decrease the probability of having or developing a disease after taking the index test. Hence, current ML models have relatively sufficient ability to predict the occurrence of hypoglycemia, especially RF algorithms with a PLR of 13.9 (95% CI 10.1-18.9) and an NLR of 0.14 (95% CI 0.08-0.22). On the contrary, although the PLR of NNM algorithms was 5.9 (95% CI 3.2-10.8), their sensitivity and NLR were 0.50 (95% CI 0.16-0.84) and 0.54 (95% CI 0.24-1.21), respectively, which is far from satisfactory. Although RF algorithms seem to be able to capture the complex, nonlinear patterns affecting hypoglycemia [56], it was still not enough to determine which algorithm shows the best

performance, as the test scenarios were quite different and there was high heterogeneity between studies.

Regarding ML models for detecting hypoglycemia, the pooled sensitivity, specificity, PLR, and NLR were 0.74 (95% CI 0.70-0.78), 0.70 (0.56-0.81), 2.4 (1.6-3.7), and 0.37 (0.29-0.46), respectively, which indicates that the algorithms generate small changes in probability [69]. Nevertheless, it does not mean that ML models combined with ECG or EEG monitoring, which we found in 13 of 17 studies, should not be further investigated. Considering patients with both DM and cardiovascular risk, or patients under intensive care and in a coma, combined ML models and ECG or EEG signals might be able to avoid deficits in physical and cognitive function and death caused by hypoglycemia [70].

Strengths and Limitations

The study has several limitations. First, although we developed a comprehensive search strategy, there was still a possibility of potential missing studies. To further increase the rate of literature retrieval, we included the main medical databases with a feasible search strategy, including PubMed, Embase, Web of Science, and IEEE Explore, and references from relevant studies were also screened for eligibility to avoid omissions. Second, statistically significant high heterogeneity was detected in all subgroups, with different sources of heterogeneity, including different types of DM, ML models, data sources, reference index, time and setting of data collection, and threshold of hypoglycemia, among studies. To address this issue, hierarchical analysis and meta-regression analysis were carried out in different subgroups to explore the possible sources of heterogeneity. Furthermore, for several studies that provided no required outcome measures or had inconsistent outcome measures, relevant estimation methods were used to calculate the indicators, which might have led to a certain amount of estimation error. However, the estimation error was small enough to be accepted owing to an appropriate estimation method, and the results of this study were further enriched. However, future studies are required to report all relevant outcome measures for further evaluation.

Future Directions

In future, more accurate ML models will be used for BG management, which will certainly improve the quality of life of patients with DM and reduce the burden of adverse BG events. First, as mentioned before, current ML models have relatively sufficient ability to predict BG levels and hypoglycemia, and the fact that an extended PH is more beneficial for increasing the time available for patients and clinicians to respond still needs to be emphasized [15]. Hence, future studies should focus on enhancing the performance of ML models in longer PHs (ie, 60 minutes). Second, most of the raw data from CGM devices are highly imbalanced due to the low incidence of adverse BG events, which may lead to several performance distortions. Previous studies have reported several approaches to reduce the data imbalance, including oversampling [71] and cost-based learning [15]. However, to the best of our knowledge, few studies have investigated the effectiveness of those approaches in BG management models, which needs to be further studied in the future. Furthermore, the high variability of BG levels in the human body due to several factors, such as meal intake, high-intensity exercise, and insulin dosage, creates challenges for ML models; thus, future works need to integrate these factors with existing models to further enhance their accuracy [22,51]. It is also necessary to consider the computational complexity and convenience of use for patients and physicians. Moreover, several studies have implied that a combination of ML models and features extracted from CGM profiles can achieve better predictability compared to an ML model alone [15,56]. Recently, studies have focused on more novel deep learning models, such as transformers, which have also been proved clinically useful [72]. Therefore, further studies that focus on optimizing the structure of an ensemble method are needed to explore more models with a new structure. Lastly, it should be mentioned that although several studies have achieved high performance using relatively small data set [29,31,32,35,39,47,57], which can reduce the difficulty in model development, it also creates a concern about whether this will decrease the generalization ability of the models. Most of the models were developed and tested with a certain data set, and few of them have been prospectively validated in a clinical setting. Therefore, they need to be applied in clinical practice and be updated, as needed, to provide real-time feedback for the automatic collection of BG levels and generate a basis for prompt medical intervention [73].

Conclusion

In summary, in predicting precise BG levels, the RMSE increases with an increase in the PH, and the NNM shows the relatively highest performance among all the ML models. Meanwhile, according to the PLR and NLR, current ML models have sufficient ability to predict adverse BG (hypoglycemia) events, while their ability to detect adverse BG events needs to be enhanced. Future studies are required to focus on improving the performance and using ML models in clinical practice [70,73].

Acknowledgments

The study was funded by the National Natural Science Foundation of China (grant no. 82073663) and the Shaanxi Provincial Research and Development Program Foundation (grant nos. 2017JM7008 and 2022SF-245).

Data Availability

The data sets used and analyzed during the study are available from the corresponding author upon reasonable request.

Authors' Contributions

YW and CC conceived and designed the study. KL and LL undertook the literature review and extracted data. KL, LL, and JJ interpreted the data. KL, YM, and SL wrote the first draft of the manuscript, with revision by YW, ZL, CP, and ZY. All authors have read and approved the final version of the manuscript and had final responsibility for submitting it for publication.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplemental plot1-forest (RMSE PH=30). PH: prediction horizon; RMSE: root mean square error.
[PNG File , 808 KB - medinform_v11i1e47833_app1.png]

Multimedia Appendix 2
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) checklist.
[PDF File (Adobe PDF File), 66 KB - medinform_v11i1e47833_app2.pdf]

Multimedia Appendix 3
Supplemental plot2-forest (RMSE PH=60). PH: prediction horizon; RMSE: root mean square error.
[PNG File , 565 KB - medinform_v11i1e47833_app3.png]

Multimedia Appendix 4
Supplemental plot3-forest (RMSE PH=15). PH: prediction horizon; RMSE: root mean square error.
[PNG File , 1014 KB - medinform_v11i1e47833_app4.png]

Multimedia Appendix 5
Supplemental plot4-forest (RMSE PH=45). PH: prediction horizon; RMSE: root mean square error.
[PNG File , 838 KB - medinform_v11i1e47833_app5.png]

Multimedia Appendix 6
Supplemental plot5 - metaregression (pre-all).
[PNG File , 130 KB - medinform_v11i1e47833_app6.png]

Multimedia Appendix 7
Supplemental plot5-metaregression(pre-NN).
[PNG File , 136 KB - medinform_v11i1e47833_app7.png]

Multimedia Appendix 8
Supplemental plot5-metaregression(pre-SVM).
[PNG File , 132 KB - medinform_v11i1e47833_app8.png]

Multimedia Appendix 9
Supplemental plot5-metaregression(det-all).
[PNG File , 129 KB - medinform_v11i1e47833_app9.png]

Multimedia Appendix 10
supplemental plot5-metaregression(det-NN).
[PNG File , 123 KB - medinform_v11i1e47833_app10.png]

Multimedia Appendix 11
Supplemental plot5-metaregression(det-SVM).
[PNG File , 132 KB - medinform_v11i1e47833_app11.png]

References

1. Oviedo S, Vehí J, Calm R, Armengol J. A review of personalized blood glucose prediction strategies for T1DM patients. Int J Numer Method Biomed Eng 2017 Jun;33(6):e2833. [doi: 10.1002/cnm.2833] [Medline: 27644067]

2. Saeedi P, Petersohn I, Salpea P, Malanda B, Karuranga S, Unwin N, IDF Diabetes Atlas Committee. Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: results from the International Diabetes Federation Diabetes Atlas, 9 edition. Diabetes Res Clin Pract 2019 Nov;157:107843. [doi: 10.1016/j.diabres.2019.107843] [Medline: 31518657]
3. BMC Medicine. Diabetes education for better personalized management in pediatric patients. BMC Med 2023 Jan 24;21(1):30 [FREE Full text] [doi: 10.1186/s12916-022-02709-2] [Medline: 36690983]
4. Chen D, Wang M, Shang X, Liu X, Liu X, Ge T, et al. Development and validation of an incidence risk prediction model for early foot ulcer in diabetes based on a high evidence systematic review and meta-analysis. Diabetes Res Clin Pract 2021 Oct;180:109040. [doi: 10.1016/j.diabres.2021.109040] [Medline: 34500005]
5. Li Y, Su X, Ye Q, Guo X, Xu B, Guan T, et al. The predictive value of diabetic retinopathy on subsequent diabetic nephropathy in patients with type 2 diabetes: a systematic review and meta-analysis of prospective studies. Ren Fail 2021 Dec;43(1):231-240 [FREE Full text] [doi: 10.1080/0886022X.2020.1866010] [Medline: 33478336]
6. Wu B, Niu Z, Hu F. Study on risk factors of peripheral neuropathy in type 2 diabetes mellitus and establishment of prediction model. Diabetes Metab J 2021 Jul;45(4):526-538 [FREE Full text] [doi: 10.4093/dmj.2020.0100] [Medline: 34352988]
7. Bellemo V, Lim G, Rim TH, Tan GSW, Cheung CY, Sadda S, et al. Artificial intelligence screening for diabetic retinopathy: the real-world emerging application. Curr Diab Rep 2019 Jul 31;19(9):72. [doi: 10.1007/s11892-019-1189-3] [Medline: 31367962]
8. Jain AMC, Ahmeti I, Bogoev M, Petrovski G, Milenkovikj T, Krstevska B, et al. A new classification of diabetic foot complications: a simple and effective teaching tool. J Diab Foot Comp 2012;4(1):1-5.
9. Okonofua FE, Odimegwu C, Ajabor H, Daru PH, Johnson A. Assessing the prevalence and determinants of unwanted pregnancy and induced abortion in Nigeria. Stud Fam Plann 1999 Mar;30(1):67-77. [doi: 10.1111/j.1728-4465.1999.00067.x] [Medline: 10216897]
10. Jin Y, Li F, Vimalananda VG, Yu H. Automatic detection of hypoglycemic events from the electronic health record notes of diabetes patients: empirical study. JMIR Med Inform 2019 Nov 08;7(4):e14340 [FREE Full text] [doi: 10.2196/14340] [Medline: 31702562]
11. Lipska KJ, Ross JS, Wang Y, Inzucchi SE, Minges K, Karter AJ, et al. National trends in US hospital admissions for hyperglycemia and hypoglycemia among Medicare beneficiaries, 1999 to 2011. JAMA Intern Med 2014 Jul;174(7):1116-1124 [FREE Full text] [doi: 10.1001/jamainternmed.2014.1824] [Medline: 24838229]
12. Zou Y, Zhao L, Zhang J, Wang Y, Wu Y, Ren H, et al. Development and internal validation of machine learning algorithms for end-stage renal disease risk prediction model of people with type 2 diabetes mellitus and diabetic kidney disease. Ren Fail 2022 Dec;44(1):562-570 [FREE Full text] [doi: 10.1080/0886022X.2022.2056053] [Medline: 35373711]
13. Felizardo V, Garcia NM, Pombo N, Megdiche I. Data-based algorithms and models using diabetics real data for blood glucose and hypoglycaemia prediction - a systematic literature review. Artif Intell Med 2021 Aug;118:102120. [doi: 10.1016/j.artmed.2021.102120] [Medline: 34412843]
14. Rodbard D. Continuous glucose monitoring: a review of recent studies demonstrating improved glycemic outcomes. Diabetes Technol Ther 2017 Jun;19(S3):S25-S37 [FREE Full text] [doi: 10.1089/dia.2017.0035] [Medline: 28585879]
15. Seo W, Lee Y, Lee S, Jin S, Park S. A machine-learning approach to predict postprandial hypoglycemia. BMC Med Inform Decis Mak 2019 Nov 06;19(1):210 [FREE Full text] [doi: 10.1186/s12911-019-0943-4] [Medline: 31694629]
16. Nguyen LB, Nguyen AV, Ling SH, Nguyen HT. Combining genetic algorithm and Levenberg-Marquardt algorithm in training neural network for hypoglycemia detection using EEG signals. Annu Int Conf IEEE Eng Med Biol Soc 2013;2013:5386-5389. [doi: 10.1109/EMBC.2013.6610766] [Medline: 24110953]
17. Rodríguez-Rodríguez I, Rodríguez J, Woo WL, Wei B, Pardo-Quiles D. A comparison of feature selection and forecasting machine learning algorithms for predicting glycaemia in type 1 diabetes mellitus. Appl Sci 2021 Feb 16;11(4):1742. [doi: 10.3390/app11041742]
18. Wang Y, Wu X, Mo X. A novel adaptive-weighted-average framework for blood glucose prediction. Diabetes Technol Ther 2013 Oct;15(10):792-801 [FREE Full text] [doi: 10.1089/dia.2013.0104] [Medline: 23883406]
19. San PP, Ling SH, Soe NN, Nguyen HT. A novel extreme learning machine for hypoglycemia detection. Annu Int Conf IEEE Eng Med Biol Soc 2014;2014:302-305. [doi: 10.1109/EMBC.2014.6943589] [Medline: 25569957]
20. Pérez-Gandía C, Facchinetti A, Sparacino G, Cobelli C, Gómez EJ, Rigla M, et al. Artificial neural network algorithm for online glucose prediction from continuous glucose monitoring. Diabetes Technol Ther 2010 Jan;12(1):81-88. [doi: 10.1089/dia.2009.0076] [Medline: 20082589]
21. Prendin F, Del Favero S, Vettoretti M, Sparacino G, Facchinetti A. Forecasting of glucose levels and hypoglycemic events: head-to-head comparison of linear and nonlinear data-driven algorithms based on continuous glucose monitoring data only. Sensors (Basel) 2021 Feb 27;21(5):1647 [FREE Full text] [doi: 10.3390/s21051647] [Medline: 33673415]
22. Zhu T, Li K, Chen J, Herrero P, Georgiou P. Dilated recurrent neural networks for glucose forecasting in type 1 diabetes. J Healthc Inform Res 2020 Sep 12;4(3):308-324 [FREE Full text] [doi: 10.1007/s41666-020-00068-2] [Medline: 35415447]
23. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: the PRISMA statement. PLoS Med 2009 Jul 21;6(7):e1000097 [FREE Full text] [doi: 10.1371/journal.pmed.1000097] [Medline: 19621072]

24. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JPA, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. BMJ 2009 Jul 21;339:b2700 [FREE Full text] [doi: 10.1136/bmj.b2700] [Medline: 19622552]
25. Akl E, Altman D, Aluko P, Askie L, Beaton D, Berlin J. Cochrane Handbook for Systematic Reviews of Interventions. New York, NY: John Wiley & Sons; 2019.
26. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, QUADAS-2 Group. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011 Oct 18;155(8):529-536 [FREE Full text] [doi: 10.7326/0003-4819-155-8-201110180-00009] [Medline: 22007046]
27. White I. Multivariate random-effects meta-regression: updates to Mvmeta. Stata J 2011 Jul 01;11(2):255-270. [doi: 10.1177/1536867x1101100206]
28. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003 Sep 06;327(7414):557-560 [FREE Full text] [doi: 10.1136/bmj.327.7414.557] [Medline: 12958120]
29. Parcerisas A, Contreras I, Delecourt A, Bertachi A, Beneyto A, Conget I, et al. A machine learning approach to minimize nocturnal hypoglycemic events in type 1 diabetic patients under multiple doses of insulin. Sensors (Basel) 2022 Feb 21;22(4):1665 [FREE Full text] [doi: 10.3390/s22041665] [Medline: 35214566]
30. Stuart K, Adderley NJ, Marshall T, Rayman G, Sitch A, Manley S, et al. Predicting inpatient hypoglycaemia in hospitalized patients with diabetes: a retrospective analysis of 9584 admissions with diabetes. Diabet Med 2017 Oct 12;34(10):1385-1391. [doi: 10.1111/dme.13409] [Medline: 28632918]
31. Bertachi A, Viñals C, Biagi L, Contreras I, Vehí J, Conget I, et al. Prediction of nocturnal hypoglycemia in adults with type 1 diabetes under multiple daily injections using continuous glucose monitoring and physical activity monitor. Sensors (Basel) 2020 Mar 19;20(6):1705 [FREE Full text] [doi: 10.3390/s20061705] [Medline: 32204318]
32. Elhadd T, Mall R, Bashir M, Palotti J, Fernandez-Luque L, Farooq F, for PROFAST-Ramadan Study Group. Artificial intelligence (AI) based machine learning models predict glucose variability and hypoglycaemia risk in patients with type 2 diabetes on a multiple drug regimen who fast during ramadan (the PROFAST - IT Ramadan study). Diabetes Res Clin Pract 2020 Nov;169:108388 [FREE Full text] [doi: 10.1016/j.diabres.2020.108388] [Medline: 32858096]
33. Mosquera-Lopez C, Dodier R, Tyler NS, Wilson LM, El Youssef J, Castle JR, et al. Predicting and preventing nocturnal hypoglycemia in type 1 diabetes using big data analytics and decision theoretic analysis. Diabetes Technol Ther 2020 Nov;22(11):801-811 [FREE Full text] [doi: 10.1089/dia.2019.0458] [Medline: 32297795]
34. Ruan Y, Bellot A, Moysova Z, Tan GD, Lumb A, Davies J, et al. Predicting the risk of inpatient hypoglycemia with machine learning using electronic health records. Diabetes Care 2020 Jul;43(7):1504-1511. [doi: 10.2337/dc19-1743] [Medline: 32350021]
35. Guemes A, Cappon G, Hernandez B, Reddy M, Oliver N, Georgiou P, et al. Predicting quality of overnight glycaemic control in type 1 diabetes using binary classifiers. IEEE J Biomed Health Inform 2020 May;24(5):1439-1446 [FREE Full text] [doi: 10.1109/JBHI.2019.2938305] [Medline: 31536025]
36. Jensen MH, Dethlefsen C, Vestergaard P, Hejlesen O. Prediction of nocturnal hypoglycemia from continuous glucose monitoring data in people with type 1 diabetes: a proof-of-concept study. J Diabetes Sci Technol 2020 Mar;14(2):250-256 [FREE Full text] [doi: 10.1177/1932296819868727] [Medline: 31390891]
37. Oviedo S, Contreras I, Quirós C, Giménez M, Conget I, Vehi J. Risk-based postprandial hypoglycemia forecasting using supervised learning. Int J Med Inform 2019 Jun;126:1-8. [doi: 10.1016/j.ijmedinf.2019.03.008] [Medline: 31029250]
38. Toffanin C, Aiello EM, Cobelli C, Magni L. Hypoglycemia prevention via personalized glucose-insulin models identified in free-living conditions. J Diabetes Sci Technol 2019 Nov;13(6):1008-1016 [FREE Full text] [doi: 10.1177/1932296819880864] [Medline: 31645119]
39. Plis K, Bunescu R, Marling C, Shubrook J, Schwartz F. A machine learning approach to predicting blood glucose levels for diabetes management. 2014 Presented at: AAAI-14: 2014 Association for the Advancement of Artificial Intelligence Workshop; 2014; Ohio.
40. Chan K, Ling S, Dillon T, Nguyen H. Diagnosis of hypoglycemic episodes using a neural network based rule discovery system. Expert Syst Appl 2011 Aug 19;38(8):9799-9808 [FREE Full text] [doi: 10.1016/j.eswa.2011.02.020] [Medline: 37860015]
41. Nguyen HT, Jones TW. Detection of nocturnal hypoglycemic episodes using EEG signals. Annu Int Conf IEEE Eng Med Biol Soc 2010;2010:4930-4933. [doi: 10.1109/IEMBS.2010.5627233] [Medline: 21096665]
42. Rubega M, Scarpa F, Teodori D, Sejling A, Frandsen CS, Sparacino G. Detection of hypoglycemia using measures of EEG complexity in type 1 diabetes patients. Entropy (Basel) 2020 Jan 09;22(1):81 [FREE Full text] [doi: 10.3390/e22010081] [Medline: 33285854]
43. Chen J, Lalor J, Liu W, Druhl E, Granillo E, Vimalananda VG, et al. Detecting hypoglycemia incidents reported in patients' secure messages: using cost-sensitive learning and oversampling to reduce data imbalance. J Med Internet Res 2019 Mar 11;21(3):e11990 [FREE Full text] [doi: 10.2196/11990] [Medline: 30855231]
44. Jensen MH, Christensen TF, Tarnow L, Seto E, Dencker Johansen M, Hejlesen OK. Real-time hypoglycemia detection from continuous glucose monitoring data of subjects with type 1 diabetes. Diabetes Technol Ther 2013 Jul;15(7):538-543. [doi: 10.1089/dia.2013.0069] [Medline: 23631608]

45. Skladnev VN, Ghevondian N, Tarnavskii S, Paramalingam N, Jones TW. Clinical evaluation of a noninvasive alarm system for nocturnal hypoglycemia. J Diabetes Sci Technol 2010 Jan 01;4(1):67-74 [FREE Full text] [doi: 10.1177/193229681000400109] [Medline: 20167169]
46. Iaione F, Marques JLB. Methodology for hypoglycaemia detection based on the processing, analysis and classification of the electroencephalogram. Med Biol Eng Comput 2005 Jul;43(4):501-507. [doi: 10.1007/BF02344732] [Medline: 16255433]
47. Bertachi A, Biagi L, Contreras I, Luo N, Vehí J. Prediction of blood glucose levels and nocturnal hypoglycemia using physiological models and artificial neural networks. 2013 Presented at: 3rd International Workshop on Knowledge Discovery in Healthcare Data; July 13, 2018; Stockholm, Sweden.
48. Eljil KAAS. Predicting Hypoglycemia in Diabetic Patients Using Machine Learning Techniques. United Arab Emirates: American University of Sharjah; 2014.
49. D'Antoni F, Merone M, Piemonte V, Iannello G, Soda P. Auto-regressive time delayed jump neural network for blood glucose levels forecasting. Knowl Based Syst 2020 Sep;203:106134. [doi: 10.1016/j.knosys.2020.106134]
50. Amar Y, Shilo S, Oron T, Amar E, Phillip M, Segal E. Clinically accurate prediction of glucose levels in patients with type 1 diabetes. Diabetes Technol Ther 2020 Aug 01;22(8):562-569. [doi: 10.1089/dia.2019.0435] [Medline: 31928415]
51. Li K, Liu C, Zhu T, Herrero P, Georgiou P. GluNet: a deep learning framework for accurate glucose forecasting. IEEE J Biomed Health Inform 2020 Feb;24(2):414-423. [doi: 10.1109/jbhi.2019.2931842]
52. Zecchin C, Facchinetti A, Sparacino G, De Nicolao G, Cobelli C. Neural network incorporating meal information improves accuracy of short-time prediction of glucose concentration. IEEE Trans Biomed Eng 2012 Jun;59(6):1550-1560. [doi: 10.1109/TBME.2012.2188893] [Medline: 22374344]
53. Mohebbi A, Johansen AR, Hansen N, Christensen PE, Tarp JM, Jensen ML, et al. Short term blood glucose prediction based on continuous glucose monitoring data. Annu Int Conf IEEE Eng Med Biol Soc 2020 Jul;2020:5140-5145. [doi: 10.1109/EMBC44109.2020.9176695] [Medline: 33019143]
54. Daniels J, Herrero P, Georgiou P. A multitask learning approach to personalized blood glucose prediction. IEEE J Biomed Health Inform 2022 Jan;26(1):436-445. [doi: 10.1109/JBHI.2021.3100558] [Medline: 34314367]
55. Alfian G, Syafrudin M, Anshari M, Benes F, Atmaji F, Fahrurrozi I, et al. Blood glucose prediction model for type 1 diabetes based on artificial neural network with time-domain features. Biocybern Biomed Eng 2020 Oct;40(4):1586-1599 [FREE Full text] [doi: 10.1016/j.bbe.2020.10.004]
56. Dave D, DeSalvo DJ, Haridas B, McKay S, Shenoy A, Koh CJ, et al. Feature-based machine learning model for real-time hypoglycemia prediction. J Diabetes Sci Technol 2021 Jul 01;15(4):842-855 [FREE Full text] [doi: 10.1177/1932296820922622] [Medline: 32476492]
57. Marcus Y, Eldor R, Yaron M, Shaklai S, Ish-Shalom M, Shefer G, et al. Improving blood glucose level predictability using machine learning. Diabetes Metab Res Rev 2020 Nov 14;36(8):e3348. [doi: 10.1002/dmrr.3348] [Medline: 32445286]
58. Reddy R, Resalat N, Wilson LM, Castle JR, El Youssef J, Jacobs PG. Prediction of hypoglycemia during aerobic exercise in adults with type 1 diabetes. J Diabetes Sci Technol 2019 Sep;13(5):919-927 [FREE Full text] [doi: 10.1177/1932296818823792] [Medline: 30650997]
59. Sampath S, Tkachenko P, Renard E, Pereverzev SV. Glycemic control indices and their aggregation in the prediction of nocturnal hypoglycemia from intermittent blood glucose measurements. J Diabetes Sci Technol 2016 Nov;10(6):1245-1250 [FREE Full text] [doi: 10.1177/1932296816670400] [Medline: 27660190]
60. Sudharsan B, Peeples M, Shomali M. Hypoglycemia prediction using machine learning models for patients with type 2 diabetes. J Diabetes Sci Technol 2015 Jan;9(1):86-90 [FREE Full text] [doi: 10.1177/1932296814554260] [Medline: 25316712]
61. Nuryani N, Ling SSH, Nguyen HT. Electrocardiographic signals and swarm-based support vector machine for hypoglycemia detection. Ann Biomed Eng 2012 Apr;40(4):934-945. [doi: 10.1007/s10439-011-0446-7] [Medline: 22012087]
62. San PP, Ling SH, Nuryani N, Nguyen H. Evolvable rough-block-based neural network and its biomedical application to hypoglycemia detection system. IEEE Trans Cybern 2014 Aug;44(8):1338-1349. [doi: 10.1109/TCYB.2013.2283296] [Medline: 24122616]
63. Ling SH, Nguyen HT. Natural occurrence of nocturnal hypoglycemia detection using hybrid particle swarm optimized fuzzy reasoning model. Artif Intell Med 2012 Jul;55(3):177-184. [doi: 10.1016/j.artmed.2012.04.003] [Medline: 22698854]
64. Ling SH, San PP, Nguyen HT. Non-invasive hypoglycemia monitoring system using extreme learning machine for type 1 diabetes. ISA Trans 2016 Sep;64:440-446. [doi: 10.1016/j.isatra.2016.05.008] [Medline: 27311357]
65. Nguyen LB, Nguyen AV, Ling SH, Nguyen HT. An adaptive strategy of classification for detecting hypoglycemia using only two EEG channels. Annu Int Conf IEEE Eng Med Biol Soc 2012;2012:3515-3518. [doi: 10.1109/EMBC.2012.6346724] [Medline: 23366685]
66. Ngo CQ, Chai R, Nguyen TV, Jones TW, Nguyen HT. Electroencephalogram spectral moments for the detection of nocturnal hypoglycemia. IEEE J Biomed Health Inform 2020 May;24(5):1237-1245. [doi: 10.1109/JBHI.2019.2931782] [Medline: 31369389]
67. Ngo CQ, Truong BCQ, Jones TW, Nguyen HT. Occipital EEG activity for the detection of nocturnal hypoglycemia. Annu Int Conf IEEE Eng Med Biol Soc 2018 Jul;2018:3862-3865. [doi: 10.1109/EMBC.2018.8513069] [Medline: 30441206]

68. Nuryani N, Ling SH, Nguyen HT. Hypoglycaemia detection for type 1 diabetic patients based on ECG parameters using fuzzy support vector machine. 2010 Presented at: IJCNN 2010: 2010 International Joint Conference on Neural Networks; July 18-23, 2010; Barcelona, Spain. [doi: 10.1109/ijcnn.2010.5596916]
69. Jaeschke R, Guyatt GH, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? The Evidence-Based Medicine Working Group. JAMA 1994 Mar 02;271(9):703-707. [doi: 10.1001/jama.271.9.703] [Medline: 8309035]
70. Kodama S, Fujihara K, Shiozaki H, Horikawa C, Yamada MH, Sato T, et al. Ability of current machine learning algorithms to predict and detect hypoglycemia in patients with diabetes mellitus: meta-analysis. JMIR Diabetes 2021 Jan 29;6(1):e22458 [FREE Full text] [doi: 10.2196/22458] [Medline: 33512324]
71. McShinsky R, Marshall B. Comparison of forecasting algorithms for type 1 diabetic glucose prediction on 30 and 60-minute prediction horizons. 2020 Presented at: KDH@ECAI 2020: 5th International Workshop on Knowledge Discovery in Healthcare Data co-located with 24th European Conference on Artificial Intelligence; August 29-30, 2020; Santiago de Compostela, Spain, and virtually.
72. Deng Y, Lu L, Aponte L, Angelidi AM, Novak V, Karniadakis GE, et al. Deep transfer learning and data augmentation improve glucose levels prediction in type 2 diabetes patients. NPJ Digit Med 2021 Jul 14;4(1):109 [FREE Full text] [doi: 10.1038/s41746-021-00480-x] [Medline: 34262114]
73. Van Calster B, Steyerberg EW, Wynants L, van Smeden M. There is no such thing as a validated prediction model. BMC Med 2023 Feb 24;21(1):70 [FREE Full text] [doi: 10.1186/s12916-023-02779-w] [Medline: 36829188]

Abbreviations

ARM: autoregression model
ARJNN: ARTiDe jump neural network
AUC: area under the curve
BG: blood glucose
CGM: continuous glucose monitoring
DM: diabetes mellitus
DRNN: dilated recurrent neural network
DT: decision tree
ECG: electrocardiograph
EEG: electroencephalograph
EHR: electronic health record
ML: machine learning
NLR: negative likelihood ratio
NNM: neural network model
PH: prediction horizon
PLR: positive likelihood ratio
QUADAS-2: Quality Assessment of Diagnostic Accuracy Studies
RF: random forest
RMSE: root mean square error
SROC: summary receiver operating characteristic
SUCRA: surface under the cumulative ranking
SVM: support vector machine
T1DM: type 1 diabetes mellitus
T2DM: type 2 diabetes mellitus
XGBoost: Extreme Gradient Boosting

Edited by C Lovis; submitted 03.04.23; peer-reviewed by C Toffanin, S Lee; comments to author 30.07.23; revised version received 21.08.23; accepted 12.10.23; published 20.11.23.

Please cite as:
Liu K, Li L, Ma Y, Jiang J, Liu Z, Ye Z, Liu S, Pu C, Chen C, Wan Y
Machine Learning Models for Blood Glucose Level Prediction in Patients With Diabetes Mellitus: Systematic Review and Network Meta-Analysis
JMIR Med Inform 2023;11:e47833
URL: https://medinform.jmir.org/2023/1/e47833
doi:10.2196/47833
PMID:37983072

Review

The Roles of Electronic Health Records for Clinical Trials in Low- and Middle-Income Countries: Scoping Review

Jiancheng Ye[1,2*], PhD; Shangzhi Xiong[3,4*], MS; Tengyi Wang[5], MS; Jingyi Li[6]; Nan Cheng[7]; Maoyi Tian[3,5], PhD; Yang Yang[8], PhD

[1]Weill Cornell Medicine, New York, NY, United States

[2]Northwestern University Feinberg School of Medicine, Chicago, IL, United States

[3]The George Institute for Global Health, Faulty of Medicine and Health, University of New South Wales, Sydney, Australia

[4]Global Health Research Centre, Duke Kunshan University, Kunshan, China

[5]School of Public Health, Harbin Medical University, Harbin, China

[6]School of Basic Medicine, Harbin Medical University, Harbin, China

[7]The First Affiliated Hospital of Harbin Medical University, Harbin, China

[8]School of Public Health, Shanghai Jiao Tong University School of Medicine, Shanghai, China

[*]these authors contributed equally

Corresponding Author:
Maoyi Tian, PhD
School of Public Health
Harbin Medical University
157 Baojian Road, Nangang District
Harbin, 150081
China
Phone: 86 1082800577
Email: mtian@georgeinstitute.org.cn

Abstract

Background: Clinical trials are a crucial element in advancing medical knowledge and developing new treatments by establishing the evidence base for safety and therapeutic efficacy. However, the success of these trials depends on various factors, including trial design, project planning, research staff training, and adequate sample size. It is also crucial to recruit participants efficiently and retain them throughout the trial to ensure timely completion.

Objective: There is an increasing interest in using electronic health records (EHRs)—a widely adopted tool in clinical practice—for clinical trials. This scoping review aims to understand the use of EHR in supporting the conduct of clinical trials in low- and middle-income countries (LMICs) and to identify its strengths and limitations.

Methods: A comprehensive search was performed using 5 databases: MEDLINE, Embase, Scopus, Cochrane Library, and the Cumulative Index to Nursing and Allied Health Literature. We followed the latest version of the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guideline to conduct this review. We included clinical trials that used EHR at any step, conducted a narrative synthesis of the included studies, and mapped the roles of EHRs into the life cycle of a clinical trial.

Results: A total of 30 studies met the inclusion criteria: 13 were randomized controlled trials, 3 were cluster randomized controlled trials, 12 were quasi-experimental studies, and 2 were feasibility pilot studies. Most of the studies addressed infectious diseases (15/30, 50%), with 80% (12/15) of them about HIV or AIDS and another 40% (12/30) focused on noncommunicable diseases. Our synthesis divided the roles of EHRs into 7 major categories: participant identification and recruitment (12/30, 40%), baseline information collection (6/30, 20%), intervention (8/30, 27%), fidelity assessment (2/30, 7%), primary outcome assessment (24/30, 80%), nonprimary outcome assessment (13/30, 43%), and extended follow-up (2/30, 7%). None of the studies used EHR for participant consent and randomization.

Conclusions: Despite the enormous potential of EHRs to increase the effectiveness and efficiency of conducting clinical trials in LMICs, challenges remain. Continued exploration of the appropriate uses of EHRs by navigating their strengths and limitations to ensure fitness for use is necessary to better understand the most optimal uses of EHRs for conducting clinical trials in LMICs.

(JMIR Med Inform 2023;11:e47052) doi:10.2196/47052

KEYWORDS

electronic health records; clinical trials; low- and middle-income countries

Introduction

Clinical trials are a crucial element in advancing medical knowledge and developing new treatments by establishing the evidence base for safety and therapeutic efficacy [1]. However, the success of these trials depends on various factors, including trial design, project planning, research staff training, and adequate sample size [2]. It is also crucial to recruit participants efficiently and retain them throughout the trial to ensure timely completion [3].

Randomized controlled trials (RCTs) are considered the gold standard for evaluating the benefits and risks of health care treatments. Despite their high level of evidence, RCTs are often time consuming and expensive and may be limited by strictly standardized research settings that can hinder the generalizability of their results [4]. One promising solution to this challenge is the use of electronic health records (EHRs) to conduct large and pragmatic trials [5]. However, the gap in health care resources between high-income countries (HICs) and low- and middle-income countries (LMICs) varies greatly [6,7]. Although HICs have made significant progress in using EHR for clinical trials [8-10], little is known about the effectiveness of similar applications in LMICs [11,12]. Understanding the progress made in LMICs and how EHR has been applied to clinical trials can provide valuable insights for promoting and improving population health [13]. Conducting clinical trials in LMICs can also provide a comprehensive evaluation of interventions in different settings beyond HICs [14].

This scoping review aims to comprehensively understand the roles of EHRs in the life cycle of clinical trials, determine how EHRs were implemented in clinical research settings, and further describe specifically how this technology should be used to support different types of clinical trials in an LMIC context.

Methods

This scoping review followed the latest version of the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) guideline for the entire review process [15].

Data Sources and Search Strategy

A comprehensive search was performed using 5 databases for articles published until the end of 2021: PubMed or MEDLINE, Embase, Scopus, Cochrane Library, and the Cumulative Index to Nursing and Allied Health Literature. We prepared the search terms using the patients, problem, or population; issue of interest or intervention; comparison, control, or comparator; outcome; and study type (PICOS) approach. As the search aimed to be as comprehensive as possible and correspond to the research questions, 3 domains including EHRs, clinical trials, and LMICs (based on the list on the World Bank definition) were used to develop the search strategy [16]. A combination of keywords and controlled vocabulary terms related to the target concepts was used. The search strategy was designed and developed by 2 authors (JY and SX) independently and confirmed with an experienced librarian. Multimedia Appendix 1 presents the search strategy.

Study Selection

Studies were included in this review if they met the following criteria: (1) clear indication of clinical trials; (2) EHR was involved in the trial conduct, including identification, recruitment, informed consent collection, implementation, outcome adjudication, and outcome verification; (3) the study was conducted in an LMIC; (4) the study was published until December 31, 2021; and (5) no language restrictions. The exclusion criteria were as follows: (1) absence of a clinical trial focus; (2) the primary research question was nonclinical (ie, cost analysis study); (3) not available in full text; (4) conference abstracts or posters; (5) nonresearch articles (ie, perspectives, commentaries, letters, and reviews); and (6) retrospective secondary data analysis in a clinical trial, for example, studies that used retrospective data for 2 groups of patients who received different treatments and compared their outcomes.

First, duplicate articles were eliminated from the retrieved articles. Then, 4 reviewers (JY, SX, TW, and YY) independently screened articles based on titles and abstracts to identify the studies that could potentially fit the research question and meet the eligibility criteria. Records were excluded if they were marked as irrelevant by 2 reviewers. For records that were kept or were difficult to decide based on the title or abstract, the full text was scrutinized. When disagreements regarding study inclusion occurred between the 2 reviewers, a third or fourth reviewer was involved in the discussion until consensus was reached.

Data Extraction

A data extraction form was developed for data extraction. For each included study, we first extracted the studies' basic information, including the first author's name, publication year, country, trial setting, trial design, target population, intervention, and outcome. Of note, for trial designs, we considered individual RCT, cluster RCT, quasi-experimental studies, and feasibility pilot studies. To determine how the studies used EHR in conducting the trials, we extracted information on the roles that EHR played at any step in each of the included studies.

Data Synthesis and Analysis

We conducted a descriptive analysis on the basic information of each included paper and conducted qualitative synthesis to analyze the roles that EHR played in conducting the trials and to identify their associated implications. In the qualitative synthesis process, we referred to an established framework from a publication in 2019 [11]. The study reviewed the current and prospective uses of EHR in clinical trials worldwide and outlined five steps in which EHR could be used: (1) patient identification

and recruitment, (2) participant consent and randomization, (3) intervention, (4) outcome assessment, and (5) extended follow-up [11]. On the basis of this framework, we first attempted to map our identified roles of EHR in clinical trials into these 5 steps, and then, we performed modifications by adding our identified new roles of EHR from the included studies. The identifications and articulations of new roles were based on research team discussions until consensus was reached (JY, SX, and YY). In addition, when available, we further synthesized text information about implications of using EHR in conducting clinical trials, by summarizing them as "strengths" and "limitations" under each role of EHR.

Quality Assessment

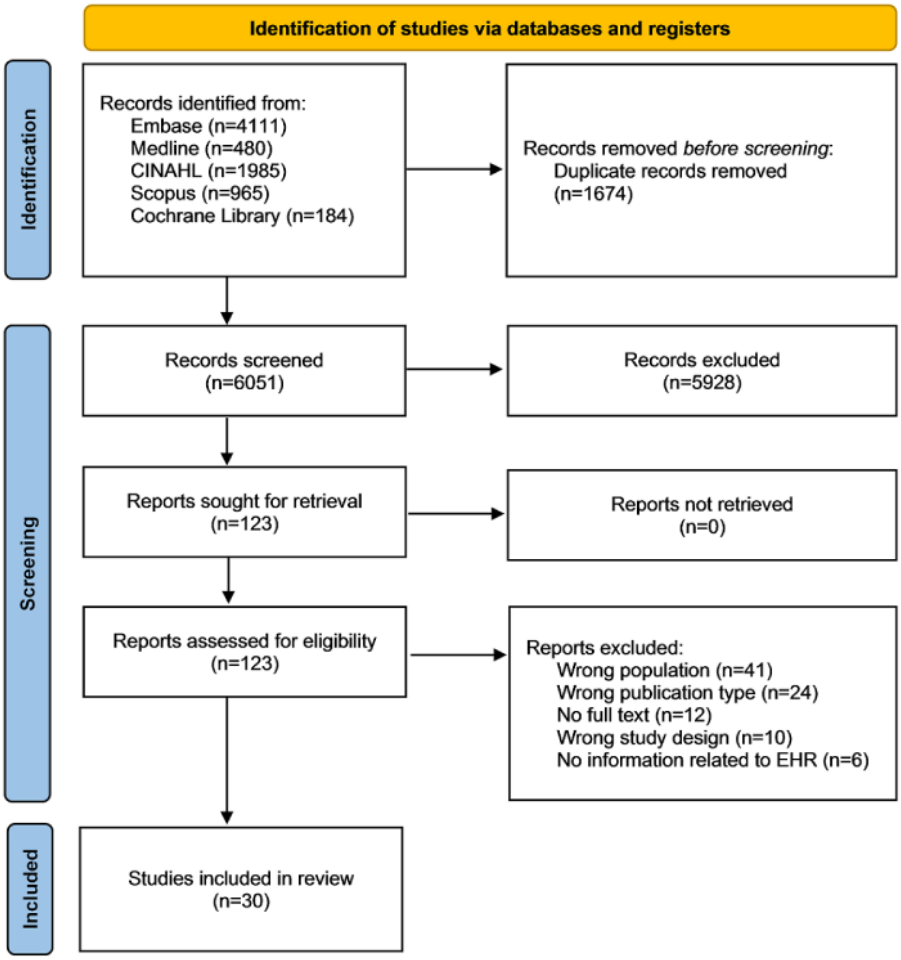
We followed the National Heart, Lung, and Blood Institute's Study Quality Assessment Tools for the quality assessment of the included studies [17]. For studies with control groups, a total of 14 questions were considered, including the adequacy of randomization, blinding of treatment assignment and outcome assessment, use of intention-to-treat analysis, and sufficiency of the sample size. For the quasi-experimental studies without independent control groups (eg, pre-post studies), a total of 12 questions were considered, including clarity in study objectives, participant eligibility criteria, prespecification of outcomes and subgroups, and sample size sufficiency. We marked studies that met ≥80% of applicable criteria as "good quality," 60% to 80% as "fair quality," and <60% as "poor quality."

Results

Selected Characteristics of the Included Studies

Figure 1 displays the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram for article selection. The initial search from the 5 databases yielded a total of 7725 references. After removal of duplicates, the first round of screening excluded 6051 references for ineligibility, leaving 123 references for full-text screening. A total of 93 references were then further excluded, primarily for wrong populations (ie, studies conducted exclusively in HICs) and wrong study types (ie, study types other than clinical trials). Finally, 30 studies were included in the data charting and analysis.

Figure 1. The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of the included studies in the review. EHR: electronic health record.



As shown in Table 1, the 30 studies were conducted in 15 LMICs, with China being the most represented (n=7, 23%), followed by Kenya (n=5, 17%). Zambia, South Africa, and Malaysia were all included in an equal number of studies, with each country being part of 3 (10%; 3 for each) studies. Of note, 27 (90%) of the 30 studies were conducted in a single LMIC, and 3 (10%) were conducted in multiple LMICs [18-20]. Table 2 presents the summaries of the characteristics of included studies. Most of the reported trials were conducted after 2010 (29/30, 97%), with the majority published in 2018 (5/30, 17%),

2019 (5/30, 17%), and 2020 (7/30, 23%). The oldest study was conducted in 2004 (1/30, 3%), whereas the most recent studies were conducted in 2021 (3/30, 10%). For study types, most included studies were RCTs (13/30, 43%), followed by quasi-experimental studies (12/30, 40%), cluster RCTs (3/30, 10%), and feasibility pilot studies (2/30, 7%). Of note, 9 quasi-experimental studies adopted a retrospective design using past data from the EHR systems. Regarding disease types, most studies addressed infectious diseases (15/30, 50%), with 40% (12/30) of them addressing HIV or AIDS. Another 12 studies focused on noncommunicable diseases, such as hypertension, diabetes, cancer, and mental illness. Three studies focused on injuries, such as hip fracture and brain injury.

Table 1. Characteristics of the included studies.

Study, year	Country	Trial design	Disease areas	Summary of interventions	Outcomes
Figar et al [21], 2004	Argentina	Quasi-experimental study	Hypertension	A complex antihypertensive intervention program for physicians of the older adults	• Systolic and diastolic blood pressure and the proportion of patients who were well controlled
Lakkis et al [22], 2011	Lebanon	Randomized controlled trial	Cancer	An informative SMS text remainder about mammogram tests	• Performance of (whether or not performed) the mammogram test
Were et al [23], 2013	Kenya	Randomized controlled trial	AIDS	A CDSS[a] with clinician-targeted computer-generated reminders	• The number of clinical visits before the completion of overdue tasks, including HIV testing, laboratory monitoring, initiating antiretroviral therapy, and making referrals
Li et al [24], 2014	China	Quasi-experimental study, using retrospective data	Cardiovascular disease	Salvianolate injection treatment over 14 days	• Indexes of liver and kidney function, including alanine aminotransferase, aspartate aminotransferase, creatinine, and blood urea nitrogen
Zhang et al [25], 2014	China	Quasi-experimental study, using retrospective data	Urologic diseases	Pharmacist interventions that included real-time monitoring of medical records and controlling the prescription of prophylactic antibiotics	• Rational use of antibiotic prophylaxis, including frequency of prophylactic antibiotic use, indications for and rate of prophylactic antibiotic use, medical cost/patient, inappropriate prophylactic antibiotic use, rate of correct antibiotic administration, and cost-benefit
Ghadieh et al [26], 2015	Lebanon	Randomized controlled trial	Bacterial infection	A set of reminders to invite participants to get the PPSV23[b] vaccine	• The vaccine administration rate in the clinics
Ali et al [18], 2016	India and Pakistan	Randomized controlled trial	Diabetes	A multicomponent quality improvement strategy	• Primary: the proportion that achieves multiple care targets • Secondary: achieving individual risk factor targets, mean risk factor changes, and patient reported outcomes (eg, health-related quality of life and treatment satisfaction scores)
Oluoch et al [27], 2015	Kenya	Cluster randomized controlled trial	AIDS	A CDSS with pop-up information and reminder whenever action is needed for an individual patient, and an alert when a patient had immunological treatment failure	• Primary: the difference between groups in the proportion of patients who experienced immunological treatment failure • Secondary: the effect of CDSS on time from detection of immunological treatment failure to clinical action and time from antiretroviral treatment initiation to first CD4[c] cell measurement
Wang et al [28], 2016	China	Quasi-experimental study, using retrospective data	Hypertension	A guideline for hypertension management	• The incidence of stroke
Al-Hashar et al [29], 2018	Oman	Randomized controlled trial	Chronic disease	Medication reconciliation on admission and discharge, medication review, bedside medication counseling, and take-home medication list	• Primary: percentage of preventable adverse drug events • Secondary: rates of readmission, rates of emergency department visits, rates of unplanned visits to hospitals or health centers, and the 3 combined

Study, year	Country	Trial design	Disease areas	Summary of interventions	Outcomes
George et al [30], 2018	Kenya	Randomized controlled trial	AIDS	A SMS intervention promoting the availability of oral self-administered HIV self-testing kits	• HIV testing rates
Mody et al [31], 2018	Zambia	Quasi-experimental study, using retrospective data	AIDS	A new HIV treatment guideline	• Timely antivirus treatment initiation, retention in care at 6 mo, and being retained and on antiretroviral treatment at 6 mo
Bachmann et al [32], 2019	Brazil	Cluster randomized controlled trial	Asthma or chronic obstructive pulmonary disease	A guide and training program for doctors and nurses	• Primary: composite scores of treatment changes, spirometry, and new asthma and COPD[d] diagnosis rates • Secondary: the disaggregated treatment and spirometry components of asthma and COPD scores, prescriptions (eg, support tobacco cessation, depression), diseases diagnosed for the first time (eg, cardiovascular disease and diabetes mellitus), and cardiovascular risk assessed
Engelbrecht et al [33], 2018	South Africa	Quasi-experimental study, using retrospective data	Mental illness	A mental health therapy in an occupational therapy-led day treatment center	• Hospitalization days, frequency of attendance and admissions to hospital, frequency of attendance and number of days spent in hospital, and attendance rate at day treatment center
Ismail et al [34], 2019	Saudi Arabia	Quasi-experimental study	Renal disease	A patient-centered pharmacist care in the hemodialysis unit, using comprehensive medication review through medication therapy management and motivational interviewing	• Primary: changes in serum phosphate levels and differences in number of medications • Secondary: systolic blood pressure, serum low-density lipoprotein levels, glycosylated hemoglobin levels, the prevalence and types of medication-related problems, and the rates of therapeutic interventions acceptances or rejections
Kelvin et al [35], 2018	Kenya	Randomized controlled trial	AIDS	An intervention consisting of sending a text message and offering a brief demonstration of the self-testing kit on the site	• HIV testing rates
Lima et al [36], 2018	Brazil	Quasi-experimental study, using retrospective data	Cancer	A guideline for hemodynamic and depth of anesthesia monitoring	• Postoperative outcomes including the use of cardiac output, central venous oxygen saturation, depth of anesthesia monitoring, intraoperative total fluid volume and colloid volumes, number of patients receiving colloids and received inotropes, rates of postoperative delirium and urinary tract infection, postoperative morbidity, and length of hospital stay
Phillips et al [37], 2020	South Africa	Randomized controlled trial	AIDS	A maternal and child health service in the antenatal clinic through cessation of breastfeeding	

Study, year	Country	Trial design	Disease areas	Summary of interventions	Outcomes
					• Primary: a composite of female participants' retention in HIV care and viral suppression preceding the long-term adherence and care engagement study visit • Secondary: current use of family planning, pregnancies since the trial, maternal hospitalizations and tuberculosis diagnoses in the past year
Tay et al [38], 2019	Malaysia	Quasi-experimental study	Bacterial infection	An education program for both physicians and patients on the rational use of antibiotics for upper respiratory infection and acute diarrhea	• Antibiotic prescription rate and rates of reattendance or hospital admission
Wu et al [39], 2019	China	Quasi-experimental study, using retrospective data	Hip fracture	A comanagement program involving both orthopedic surgeons and geriatricians embedding in a pathway of care spanning emergency department presentation to discharge from hospital	• Primary: the proportion of patients who received surgery within 48 h of admission to a ward • Secondary: the proportion of patients who were admitted to a ward within 4 h of presentation to emergency department, who developed a pressure ulcer, who received geriatrician care, and who received osteoporosis and falls prevention assessment
Ali et al [40], 2020	India	Randomized controlled trial	Depression and diabetes	A multicomponent quality improvement strategy with nonphysician care coordinators and decision support EHRs[e]	• Primary: between-group difference in Symptom Checklist Depression Scale scores and a reduction in glycosylated hemoglobin, systolic blood pressure, or low-density lipoprotein • Secondary: percentage of patients who met treatment targets or had improvements in individual outcomes; percentage of patients who met all glycosylated hemoglobin, systolic blood pressure, or low-density lipoprotein targets; and mean reductions in Symptom Checklist Depression Scale scores and Patient Health Questionnaire-9 scores
Yang et al [41], 2020	China	Randomized controlled trial	Orthognathic disease	A hydroactive dressing on the nasal ala of patients undergoing orthognathic surgery	• The incidence of nasal ala pressure injury associated with nasotracheal intubation
Puttkammer et al [42], 2020	Haiti	Feasibility pilot study	AIDS	An EHR-based alert for adherence intervention	• HIV viral load status, antiretroviral treatment adherence, and proportion of patients who were never >7 d late for an antiretroviral treatment refill pickup
Roy et al [43], 2020	Zambia	Cluster randomized controlled trial	AIDS	Adherence club group intervention to improve on-time drug pickup and retention in HIV care through off-hours facility access and pharmacist-led group drug distribution	• Primary: time to first late drug pickup • Secondary: medication possession ratio, implementation outcomes (adoption, acceptability, appropriateness, feasibility, and fidelity), and viral load suppression at 12 mo

Study, year	Country	Trial design	Disease areas	Summary of interventions	Outcomes
Seth Kalichman et al [44], 2020	South Africa	Randomized controlled trial	Sexually transmitted infections	Two counseling sessions for brief risk reduction sexual behavior change and brief enhanced partner notification	• Primary: return sexually transmitted infection visits after counseling and rate of scale representing the percentage of times condoms were used • Secondary: sexually transmitted infection risk and prevention-related knowledge assessed by 4 heterogeneous items, HIV stigma assessed by the HIV Stigma Scale, and prevention skills self-efficacy
Mahmood et al [19], 2020	United Kingdom and Malaysia	Randomized controlled trial	Traumatic brain injury	A tranexamic acid treatment	• Primary: the volume of intraparenchymal hemorrhage after randomization • Secondary: presence of progressive intracranial hemorrhage, new intracranial hemorrhage, presence of cerebral infarction, composite poor outcome, and intracranial hemorrhage volume
Mody et al [45], 2021	Zambia	Quasi-experimental study, using retrospective data	AIDS	A new national HIV treatment	• Primary: rates of both antiretroviral treatment initiation and retention on antiretroviral treatment across subgroups in 3 timestamps, antiretroviral treatment initiation within 1 mo of enrollment, and retention in care on antiretroviral treatment at 12 mo
Semeere et al [20], 2021	Uganda and Kenya	Feasibility pilot study	AIDS	A rapid case ascertainment measurement shortly after diagnosis	• Performance of (whether or not performed) rapid case ascertainment
Xu et al [46], 2021	China	Quasi-experimental study, using retrospective data	Bacterial infection	A pharmacist-led intravenous to oral antibiotic conversion practice with computerized reminders	• Primary: the proportion of patients who converted to oral therapy on the day patients were eligible for the conversion • Secondary: length of IV[f] antibiotic therapy days, total length of antibiotic therapy days, and length of hospital stays
Abdulrahman et al [47], 2017	Malaysia	Randomized controlled trial	AIDS	Mobile phone reminders based on SMS, telephone call, and peer counseling	• Primary: improved scheduled clinic attendance and medication adherence self-report • Secondary: immunological, virological, and clinical measurements

[a]CDSS: clinical decision support system.

[b]PPSV23: Pneumococcal Polysaccharide Vaccine.

[c]CD4: cluster of differentiation 4.

[d]COPD: chronic obstructive pulmonary disease.

[e]EHR: electronic health record.

[f]IV: intravenous therapy.

Table 2. Summaries of the characteristics of the included studies (N=30).

Characteristics	Studies, n (%)
Year of publication	
2017 or before	10 (33)
2018-2019	10 (33)
2020-2021	10 (33)
Trial design	
Randomized controlled trials	13 (43)
Cluster randomized controlled trials	3 (10)
Quasi-experimental studies	12 (40)
Feasibility pilot study	2 (7)
Trial settings	
Primary health care	8 (27)
Secondary health care	3 (10)
Tertiary health care	8 (27)
Roles of EHR[a]	
Identification and recruitment	12 (40)
Participant consent and randomization	0 (0)
Baseline information collection	6 (20)
Intervention	8 (27)
Fidelity assessment	2 (7)
Primary outcome assessment	24 (80)
Secondary outcome assessment	13 (43)
Disease areas	
Infectious (AIDS) and maternal and perinatal conditions	15 (50)
Noncommunicable diseases (eg, diabetes, hypertension, cancer, and mental illness)	12 (40)
Injury	3 (10)

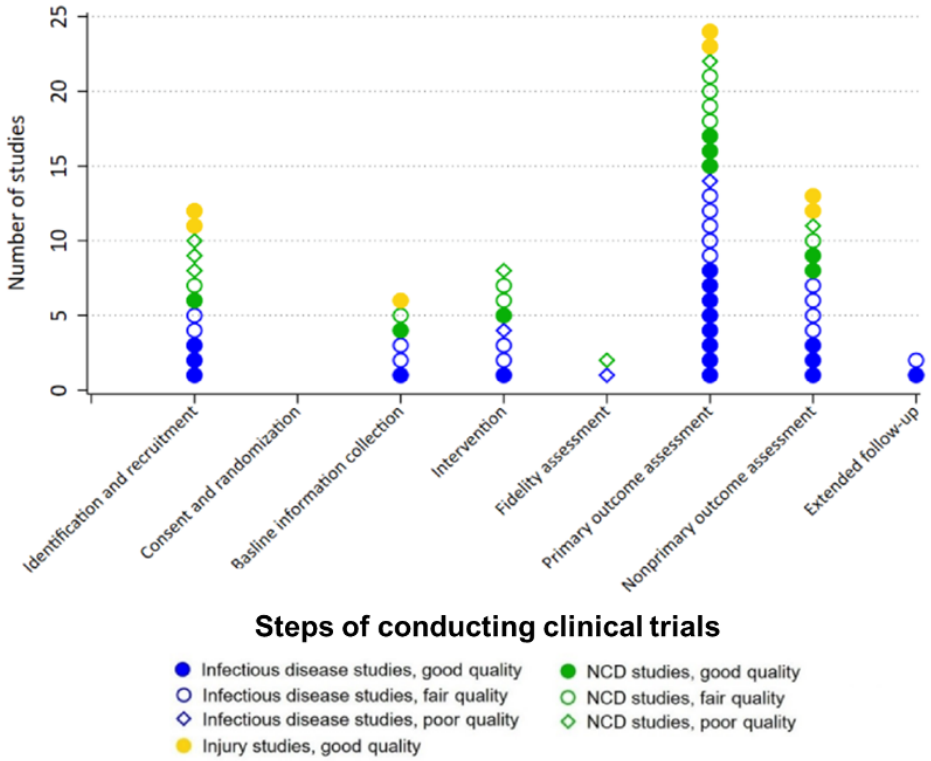
[a]EHR: electronic health record.

Interventions and Outcomes in Included Trials

The included clinical trials covered various types of interventions, ranging from single behavioral interventions such as informative SMSs for clinical appointments or vaccinations and clinical therapies such as tranexamic acid treatment or Salvianolate injection to complex multifaceted intervention packages for service quality improvements or education programs [28,31,36,45]. Corresponding to the diversity of interventions, the included studies also focused on a wide range of outcomes, including individual clinical outcomes such as blood pressure and incidence of stroke, individual behavioral outcomes such as medication use and retention in antiretroviral care, and facility-level administration data such as the number of hospital visits and vaccine administration rates [48,49].

Of the 8 steps of conducting clinical trials (Figure 2), we identified 7 in which EHR was used in the 30 included studies, with the only exception for "participant consent and randomizations." The role of EHR in primary outcome assessment (24/30, 80%) was the most commonly documented, followed by the use of EHR to assess nonprimary outcomes (13/30, 43%). For example, 1 study focused on the effects of adherence club groups on the on-time antiretroviral drug pickup among people with HIV or AIDS, whose primary outcome was "time to first late drug pickup," and secondary outcome was "the proportion of time that a patient has antiretroviral drug in their possession over 12 months," and both of them were ascertained using EHR data [43]. One study used EHR to obtain safety outcomes as a nonprimary outcome, which was the adverse change in participants' bioindicators after the intervention [24]. We summarized that there were 3 main types of information collected from the EHR: people's clinical information (eg, systolic and diastolic blood pressure), behavioral information (eg, medication use), and health facility administration data (eg, frequency of hospital visits).

Figure 2. The roles of electronic medical records in the different steps of conducting clinical trials among the included studies. NCD: noncommunicable disease.



Roles of EHRs in Trial Steps

Table 3 demonstrates the roles of electronic medical records in the different steps of conducting clinical trials among included studies. A total of 12 trials used EHR to identify and recruit study participants based on automatically or manually extracted data. These studies queried a series of data entries of clinical information from the databases to identify their target populations and retrieved the contact information to approach the potential patient participants. For example, Bachmann et al [32] identified eligible participants with asthma or chronic obstructive pulmonary disease using the International Classification of Disease diagnostic codes in a consolidated municipal EHR database in Brazil. Lakkis et al [22] used EHR to extract cell phone numbers of female participants who were recommended to undergo a screening mammogram for breast cancer.

Table 3. The roles of electronic medical records in the different steps of conducting clinical trials among included studies.

Roles of EHR[a] (number of studies)	Descriptions	Implications	References
Participant identification and Recruitment (n=12)	People's demographic, clinical, and contact information from the EHR systems was either automatically or manually extracted to help researchers identify and then recruit trial participants.	• Strength: EHR enabled researchers to access participants that were otherwise unfeasible to access. • Strength: EHR enabled researchers to identify and include all eligible patients whose interaction with health systems were electronically recorded, which increased generalizability and sample size at minimal cost. • Limitation: some eligible participants may not be documented in the EHR systems. • Limitation: artifacts, errors, and misclassifications in EHR may cause inclusion of ineligible participants or exclusion of eligible ones.	[19,21,22,24-28,30,32,35,39]
Baseline information collection (n=6)	For people who were enrolled, many trials used EHR systems to collect their baseline basic characteristics, including demographical and socioeconomical information, which were used as covariates in later analyses.	• Not mentioned.	[31,32,34,41,43,45]
Intervention (n=8)	Many trials incorporated EHR systems in the intervention packages, including three scenarios: (1) electronic reminders or alerts to health providers based on EHR data, (2) clinical decision support systems built in the EHR system, and (3) monitoring of EHR data by health providers to make clinical decisions.	• Strength: EHR could integrate different intervention components of complex interventions. • Strength: the EHR-based data review by health providers allowed real-time monitoring of patients' health. • Limitation: lack of complete and high-quality data in EHR systems would prohibit relevant, timely, and accurate clinical decision support. • Limitation: technical glitches such as server breakdown compromised the continuity of EHR-based interventions.	[18,21,23,25,40,42,46]
Fidelity (n=2)	Some trials used EHR systems to assess the fidelity of trial conduct according to predefined protocols, such as providers' adherence to study protocols and patients' adherence to lifestyle changes.	• Not mentioned.	[21,42]
Primary outcome assessment (n=24)	Most included trials used EHR systems to assess studies' primary outcomes, including clinical data (eg, blood pressure), behavioral data (eg, medication use), and service use data (frequency of hospital visits).	• Strength: using EHR for data extraction is time saving compared with conventional individual chart review. • Limitation: EHR's data quality could be questionable, subject to inaccuracy, misclassification, and incompleteness. • Limitation: some relevant information was not available in the EHR systems. • Limitation: lack of integration across or linkage with different EHR systems might cause missing information or underdetection of events. • Limitation: EHR was not able to provide information for individuals who moved out of the region in the middle of the trial. • Limitation: some data might be missing or omitted in the transformation from handwritten medical records to EHR.	[19,23-27,29-40,42-45,47,50]

Roles of EHR[a] (number of studies)	Descriptions	Implications	References
Nonprimary outcome assessment (n=13)	Many trials used EHR systems to assess studies' nonprimary outcomes, including clinical, behavioral, and service use data. In addition, 1 study used EHR to assess safety outcomes (eg, the occurrence of adverse events).	• Strength: using EHR for data extraction is time saving compared with conventional individual chart review. • Limitation: some relevant information was not available in the EHR systems, which might omit potential confounders.	[19,23,24,27,35,37-44]
Extended follow-up (n=2)	A few trials used EHR systems to follow-up with participants beyond the study time frame to determine the sustainability of effects.	• Not mentioned.	[23,37]

[a]EHR: electronic health record.

We found 8 studies that incorporated EHR into their intervention packages. There were 3 types of interventions that used EHR. First, 3 studies used electronic reminders or alerts to health care providers based on EHR data [21,42,46]. For example, Puttkammer et al [42] used the EHR systems to alert physicians of patients at elevated risk of treatment failure through automated calculations of patients' risk score based on their past EHR data. Second, 4 studies incorporated clinical decision support systems into the EHR systems [18,23,27,40]. Third, 2 studies involved manual monitoring and review of EHR data by health care providers [25,40]. For example, Ali et al [40] included both an EHR-based clinical decision support system and manual monitoring of EHR data. Their EHR-based clinical decision support system integrated patient characteristics, depressive symptom scores, and laboratory data to provide evidence-based guidelines to physicians based on treatment guidelines, and the study team also manually monitored and reviewed the EHR data and developed consensus recommendations for patients with severe symptoms.

A total of 6 studies used EHR to collect baseline information, which mainly included individual demographic, socioeconomic, and clinical information to determine the basic characteristics of the participants. For example, 1 study collected baseline information through EHRs, including age, sex, pregnancy status, HIV clinic enrollment date, and antiretroviral treatment initiation date among people with HIV or AIDS in Zambia [45].

Only 2 studies used EHR for fidelity assessment of the trial. Puttkammer et al [42] used EHR to determine health workers' "engagement" with the EHR-based computerized alert, which was defined as the frequency of the health workers' clicking on the alert to bring up the "pop up" window. Figar et al [21] used EHR to determine adherence to lifestyle changes in older patients with hypertension [21]. Both studies were feasibility studies and both lacked randomization.

Two studies used EHR to follow-up with participants beyond the study timeframe to determine the sustainability of the effects [23,37]. Phillips et al [37] used EHR to follow-up with female participants who attended a past trial on HIV or AIDS to determine the continued effects of the interventions on female participants' retention in HIV care and viral suppression. Were et al [23] queried the EHR data 3 months after the study closure to capture the study's sustained effects on the quality of pediatric HIV care in a resource-limited setting in Kenya.

Some studies discussed the strengths and limitations of using EHR in conducting clinical trials. For participant identification and recruitment, Semeere et al [20] reported that EHR enabled researchers to access participants who were otherwise unfeasible to access, and Bachmann et al [32] mentioned that EHR enabled researchers to identify and include all eligible patients whose interactions with health systems were electronically recorded, which increased generalizability and sample size at minimal cost. However, a few studies shared concerns about the inclusion of ineligible participants or exclusion of eligible participants owing to artifacts, errors, and misclassifications in EHR [22,32,35].

For studies that incorporated EHR in their interventions, Ali et al [40] mentioned the advantage of EHR to "integrate different intervention components of complex interventions," and Zhang et al [25] mentioned that EHR systems enabled real-time monitoring and reviewing of patients' health data by health providers. For limitations, by contrast, Were et al [23] argued that the lack of complete and high-quality data in EHR systems prohibits relevant, timely, and accurate clinical decision support. The server breakdown in the study by Puttkammer et al [42] represented a general concern about potential technical glitches of EHR, which could lead to risks of discontinuity of interventions.

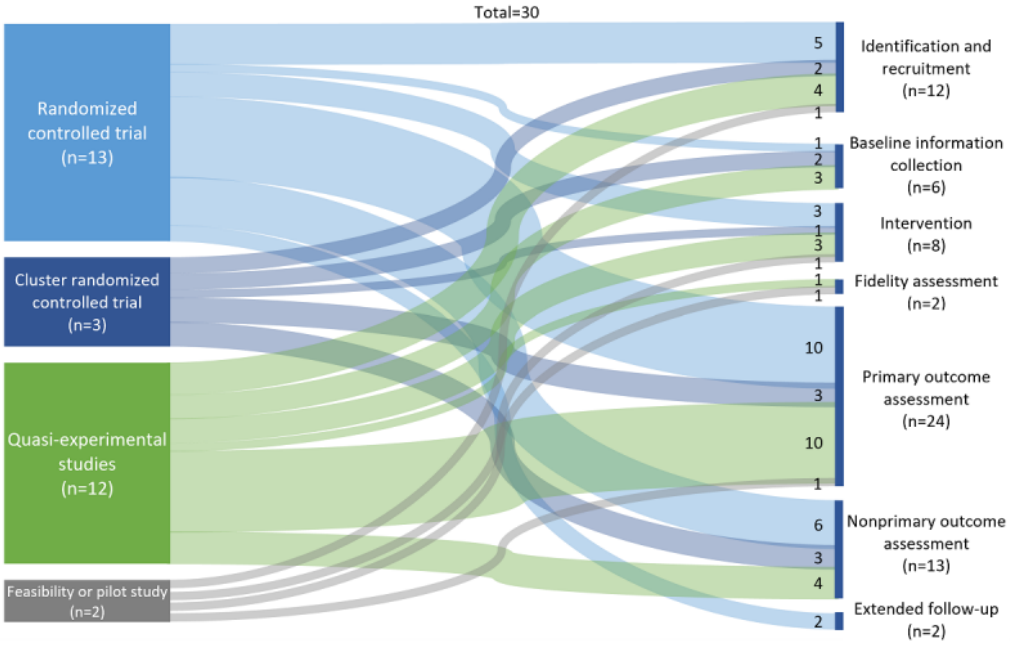
For studies that used EHR for outcome assessment, an important strength was that it was more time saving than conventional outcome research approaches, for which databases and patient registries are often fragmented and limited in the number of patients [26]. However, numerous limitations were documented, particularly concerning data quality. These limitations included inaccuracies, misclassifications, and incompleteness, which were evident in various aspects, including the use of EHR for participant identification, recruitment, baseline information collection, and interventions [35,37,45]. Data quality issues were emphasized in the study by Oluoch [26], which observed a loss of information in the transformation process from handwritten records to EHR. However, 1 study argued that these intrinsic flaws in EHR data represented the situation of real-world care delivery and was thus valuable in its own way [45]. Four studies mentioned that not all relevant information

was available in EHR systems, which could prohibit suboptimal data analysis such as underadjustment for study confounders [36,37,42,45]. Finally, some studies mentioned that the lack of integrated EHR systems across different health facilities might cause missing information or underdetected events [37,44,46].

Roles of EHRs in Trial Designs

We further found that the roles of EHR in conducting clinical trials differed by different trial designs (Figure 3). Of the 4 types of trials, RCT covered the most steps of conducting trials that used EHR and was the only type that used EHR for "extended follow-up" [23,37]. Other than that, RCTs, cluster RCTs, and quasi-experimental studies covered the same 5 steps of conducting trials using EHR: participant identification and recruitment, baseline information collection, intervention, primary and nonprimary outcome assessment. Notably, a controlled trial without randomization [21] and a feasibility study [42] were the only 2 studies that used EHR to assess the fidelity of conducting their interventions.

Figure 3. The roles of electronic health records in conducting clinical trials by different trial designs.



Quality Assessment

The quality assessment results are presented in Multimedia Appendices 2 and 3. Less than half of the included studies were found to be of good quality (14/30, 47%), 40% (12/30) of fair quality, and 13% (4/30) of poor quality. For the 20 studies that had control groups (ie, all the RCTs, cluster RCTs, non-RCTs, some quasi-experimental studies, and 1 of the feasibility pilot studies), the most common factors that compromised the quality of the studies was the lack of prespecification of outcomes and subgroups (9/30, 30%), followed by the lack of similar baseline characteristics between groups (8/30, 27%), lack of blinding in treatment assignment (7/30, 23%) and outcome assessment (7/30, 23%), and nonreport of calculations for sufficient sample size (7/30, 23%). Of the 10 studies that did not have control groups, 7 (70%) did not have interrupted time series design, 6 (60%) did not report sufficient sample size, and 5 (50%) had suboptimal representativeness of the participants to the target population.

Discussion

Summary of Principal Findings

We synthesized our findings on the use of EHR for conducting clinical trials in LMICs into a framework that depicts the typical life cycle of a clinical trial. The EHRs were primarily used for eligible participant identification or recruitment, trial outcome assessment, and intervention implementation in LMICs. The limited use of EHR was documented for participant consent, randomization, and fidelity assessment. An encouraging observation was the diversity of disease types covered in the selected studies, highlighting that EHRs have a wide appeal across various medical domains. Notably, a wide range of outcomes were assessed using EHRs in these trials, including clinical outcomes [51,52], behavioral outcomes [53,54], and health service outcomes [55].

Characteristics of EHR-Supported Trials in LMICs

The identified EHR-supported trials in LMICs were heterogeneous in terms of their targeted populations and outcomes. LMICs include, by nature, a diverse group of regions with varying population characteristics including health conditions and socioeconomic status. In general, many LMICs reportedly face challenges in terms of technological infrastructure [56], data quality, and interoperability of EHR systems, which can impact the feasibility and reliability of EHR-based trials [57]. Moreover, regulatory frameworks and guidelines for EHR-supported clinical trials have been poorly defined in some LMICs, especially for data use and security in these settings [58].

In our review, although there was a clear use of EHR in conducting clinical trials in LMICs, few of them focused on

medication. One possible explanation is that the focus of the medication-related trials is more explanatory (ie, understanding whether an intervention or medication is efficacious in an ideal setting) rather than pragmatic (ie, determining the effectiveness of interventions in real-world settings). As EHRs are usually routine health information systems rather than tools specialized for clinical trials, they are naturally more suitable for studies focusing on real-world effectiveness and implementation of an intervention but not necessarily for those focusing on intervention efficacy under strictly controlled conditions. In addition, the reported data quality issues and lack of population coverage in EHRs may also have limited their usability in efficacious studies. However, EHR can still be used to inform planning, participant recruitment, baseline statistics, and outcome extraction in medication-related trials [43]. Given that current medication development studies are primarily conducted in HICs [59], future uses of EHR may have the potential to enable more medication-related trials to be conducted in LMICs, thus increasing the representation of LMICs in study populations and geographic settings at a relatively low cost [60,61].

Challenges of Using EHRs in Clinical Trials in LMICs

There are 2 potential challenges to using EHRs in clinical trials. The first relates to possible barriers, including difficulties in accessing relevant data, linking different data sources, high financial costs, and limited familiarity with such systems [62]. The second pertains to the underreporting or exclusion of EHR information. For example, researchers may neglect the inclusion of EHR data owing to time constraints or competing priorities [63] or selectively report the EHR data that aligns with their hypotheses or desired outcomes, potentially introducing bias into the results [64]. Integrating EHR data into a clinical trial can be technically complex and time consuming [65]. Researchers may lack the resources or technical expertise to effectively integrate EHR data with the clinical trials' data set, leading to the decision to exclude or underreport it [12,66].

Another prominent observation was that almost all the included articles expressed challenges with using EHR, with some being explicit experiences and lessons. Common challenges were typically related to data availability, data quality, data interoperability, and missing data. For data availability, it meant that some relevant components to define a clinical entity were absent [37]. For data quality, the most common concern was data missingness, such as missing laboratory values in an EHR system; data artifacts were also a main concern, especially when the data were manually entered into the EHR [24]. Ultimately, the specifics of these challenges can be potentially beneficial for developing guidance on optimal EHR uses [67]. Specific to the LMICs, a tailed framework for using EHRs in clinical trials may be useful to assess the fitness of EHR for the trials [68]; using the insights from these identified challenges may be useful in ensuring the EHR selected best fits the desired need [69,70].

None of the selected studies used EHRs to collect participants' consent information and conduct randomization, which has been successfully performed in HICs. For example, the *Join Us* initiative in Australia uses the linkage of the routinely collected data including EHRs to recruit residents and collect their consent to enroll in potential clinical trials [71]. To do this in LMICs, it may be necessary to establish an updated regulatory framework for research ethics, such as the consent process for using and sharing routinely collected data and for intervention implementation. Nevertheless, obtaining informed consent for automated trials conducted using EHRs may be difficult [5]. For example, there is still debate on whether informed consent needs to be acquired when only variations of usual care are explored [72]. There are also disconnections between clinical trials that use EHRs and regulation guidelines created for traditional RCTs without the involvement of EHRs (eg, the lack of standardized requirements of institutional review boards for the use of EHRs in trials). With more examples of EHR-supported trials emerging, further research and constructive dialogues among all stakeholders are needed to alter and align the ethical norms and regulatory processes to enable more successful and accountable uses of EHRs in clinical trials in LMICs.

Notably, the quality assessment of the included trials indicated that a substantial portion of the studies did not meet the criteria for good quality. The lack of prespecification of outcomes and subgroups, for example, emerged as a notable issue, which may lead to outcome reporting bias and ambiguity in result interpretation. Other major issues included the absence of comparable baseline characteristics between the treatment and control groups, inadequate blinding, lack of considerations for sample size sufficiency, and lack of control groups. On the one hand, these commonly identified shortcomings revealed the exploratory nature of many of the included EHR-supported trials in LMICs, which needs to be addressed in future efforts to enhance rigor and credibility. On the other hand, they also implied that quasi-experimental designs, such as interrupted time series and self-controlled studies, might be the "comfort zone" for using EHRs to support future clinical trials.

Limitations

This scoping review has some limitations. First, we only focused on EHRs in clinical trials in the context of LMICs, rather than comparing the results between HICs and LMICs. Different LMICs follow different data schemas and regulatory structures, which may lead to challenges when considering generalizability. Second, we required the mention or self-tagging of EHR within the articles. This requirement likely led to a swath of missed potential articles. However, we included a wide range of synonyms of EHR in the search syntax, which should have helped address this limitation. Third, we did not test any hypothesis regarding the effect of using EHR in clinical trials, and we did not we assess the impact of using EHRs on health outcomes. Although we extracted a few characteristics that could point to the methodological quality of the studies, including the evaluation of risk of bias, we did not evaluate the intervention effects reported in the trials but merely offered a description of EHRs' roles in the trial conduct.

Conclusions

We mapped the roles of EHRs in clinical trials from the selected studies to the life cycle of clinical trials and identified opportunities to enhance the use of EHRs for clinical trials in LMICs. Specifically, the most commonly documented use was the incorporation of EHRs into clinical trials for outcome

assessment, whereas the use of EHR in collecting participant consent and conducting randomization was scarce. Efforts should be made to improve the curation of EHR data to improve data quality, explore the integration of automated processes in EHR to obtain people's consent for data use in research, and standardize regulatory frameworks for using EHR for research. Future research and practices are recommended to navigate the strengths of EHRs, such as time sensitivity and low costs, and mitigate the current challenges, such as suboptimal data quality and limited population coverage, to ensure better use of EHR in future clinical trials in LMICs. With the ongoing digitization of health information systems globally, researchers, practitioners, and policy makers are recommended to maintain continued evaluations of the availability and quality of EHRs to better understand their optimal use and unlock the full potential of EHRs for health care services and research purposes.

Acknowledgments

MT and YY are co-corresponding authors on this article. This study was financially supported by the Startup Fund for Young Faculty at Shanghai Jiao Tong University (grant BJ1-3000-22-0066), and the Science and Technology Committee of Shanghai Municipality (grant 23ZR1436400 to YY). The authors acknowledge the library support in developing the search strategy from the University of New South Wales.

Data Availability

The authors declare that all the data included in this study are available in the paper and the multimedia appendices.

Authors' Contributions

JY, SX, TW, MT, and YY conceived the study. JY, SX, TW, and YY developed the search strategy and screened the literature. JY, SX, TW, JL, NC, and YY performed the data extraction. JY and SX drafted the manuscript. All authors contributed to the interpretation of the results and revision of the manuscript. All the authors read and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search syntax by each database.
[PDF File (Adobe PDF File), 16 KB - medinform_v11i1e47052_app1.pdf]

Multimedia Appendix 2
Quality assessment of included trials with independent control groups (part A).
[PDF File (Adobe PDF File), 45 KB - medinform_v11i1e47052_app2.pdf]

Multimedia Appendix 3
Quality assessment of included trials without independent control groups (part A).
[PDF File (Adobe PDF File), 26 KB - medinform_v11i1e47052_app3.pdf]

Multimedia Appendix 4
PRISMA-ScR-Checklist.
[PDF File (Adobe PDF File), 555 KB - medinform_v11i1e47052_app4.pdf]

References

1. Turner JR, Hoofwijk TJ. Clinical trials in new drug development. J Clin Hypertens (Greenwich) 2013 May;15(5):306-309 [FREE Full text] [doi: 10.1111/jch.12085] [Medline: 23614843]
2. Lai YS, Afseth JD. A review of the impact of utilising electronic medical records for clinical research recruitment. Clin Trials 2019 Apr;16(2):194-203. [doi: 10.1177/1740774519829709] [Medline: 30764659]
3. Brueton VC, Tierney JF, Stenning S, Meredith S, Harding S, Nazareth I, et al. Strategies to improve retention in randomised trials: a Cochrane systematic review and meta-analysis. BMJ Open 2014 Feb 04;4(2):e003821 [FREE Full text] [doi: 10.1136/bmjopen-2013-003821] [Medline: 24496696]
4. Bothwell LE, Greene JA, Podolsky SH, Jones DS. Assessing the gold standard--lessons from the history of RCTs. N Engl J Med 2016 Jun 02;374(22):2175-2181. [doi: 10.1056/NEJMms1604593] [Medline: 27248626]
5. Mc Cord KA, Al-Shahi Salman R, Treweek S, Gardner H, Strech D, Whiteley W, et al. Routinely collected data for randomized trials: promises, barriers, and implications. Trials 2018 Jan 11;19(1):29 [FREE Full text] [doi: 10.1186/s13063-017-2394-5] [Medline: 29325575]

6. Kimura M, Croll P, Li B, Wong CP, Gogia S, Faud A, et al. Survey on medical records and EHR in Asia-Pacific region: languages, purposes, IDs and regulations. Methods Inf Med 2011;50(4):386-391. [doi: 10.3414/ME11-02-0002] [Medline: 21792467]
7. Ojji D, Baldridge AS, Orji IA, Shedul GL, Ojo TM, Ye J, et al. Characteristics, treatment, and control of hypertension in public primary healthcare centers in Nigeria: baseline results from the Hypertension Treatment in Nigeria Program. J Hypertens 2022 May 01;40(5):888-896 [FREE Full text] [doi: 10.1097/HJH.0000000000003089] [Medline: 35034080]
8. Ye J, Ma Q. The effects and patterns among mobile health, social determinants, and physical activity: a nationally representative cross-sectional study. AMIA Jt Summits Transl Sci Proc 2021 May 21;2021:653-662 [FREE Full text] [Medline: 34457181]
9. Ye J, Hai J, Song J, Wang Z. Multimodal data hybrid fusion and natural language processing for clinical prediction models. medRxiv. Preprint posted online August 25, 2023 [FREE Full text] [doi: 10.1101/2023.08.24.23294597]
10. Ye J, Sanchez-Pinto LN. Three data-driven phenotypes of multiple organ dysfunction syndrome preserved from early childhood to middle adulthood. AMIA Annu Symp Proc 2020 Jan 25;2020:1345-1353 [FREE Full text] [Medline: 33936511]
11. Mc Cord KA, Hemkens LG. Using electronic health records for clinical trials: where do we stand and where can we go? CMAJ 2019 Feb 04;191(5):E128-E133 [FREE Full text] [doi: 10.1503/cmaj.180841] [Medline: 30718337]
12. Rogers JR, Lee J, Zhou Z, Cheung YK, Hripcsak G, Weng C. Contemporary use of real-world data for clinical trial conduct in the United States: a scoping review. J Am Med Inform Assoc 2021 Jan 15;28(1):144-154 [FREE Full text] [doi: 10.1093/jamia/ocaa224] [Medline: 33164065]
13. Ye J, Wang Z, Hai J. Social networking service, patient-generated health data, and population health informatics: national cross-sectional study of patterns and implications of leveraging digital technologies to support mental health and well-being. J Med Internet Res 2022 Apr 29;24(4):e30898 [FREE Full text] [doi: 10.2196/30898] [Medline: 35486428]
14. Ye J. The role of health technology and informatics in a global public health emergency: practices and implications from the COVID-19 pandemic. JMIR Med Inform 2020 Jul 14;8(7):e19866 [FREE Full text] [doi: 10.2196/19866] [Medline: 32568725]
15. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
16. World Bank country and lending groups. The World Bank. URL: https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups [accessed 2023-10-16]
17. Study quality assessment tools. National Institutes of Health National Heart, Lung, and Blood Institute. URL: https://www.nhlbi.nih.gov/health-topics/study-quality-assessment-tools [accessed 2023-10-16]
18. Ali MK, Singh K, Kondal D, Devarajan R, Patel SA, Shivashankar R, et al. Effectiveness of a multicomponent quality improvement strategy to improve achievement of diabetes care goals: a randomized, controlled trial. Ann Intern Med 2016 Sep 20;165(6):399-408 [FREE Full text] [doi: 10.7326/M15-2807] [Medline: 27398874]
19. Mahmood A, Needham K, Shakur-Still H, Harris T, Jamaluddin SF, Davies D, et al. Effect of tranexamic acid on intracranial haemorrhage and infarction in patients with traumatic brain injury: a pre-planned substudy in a sample of CRASH-3 trial patients. Emerg Med J 2021 Apr;38(4):270-278 [FREE Full text] [doi: 10.1136/emermed-2020-210424] [Medline: 33262252]
20. Semeere A, Byakwaga H, Laker-Oketta M, Freeman E, Busakhala N, Wenger M, et al. Feasibility of rapid case ascertainment for cancer in East Africa: an investigation of community-representative Kaposi Sarcoma in the era of antiretroviral therapy. Cancer Epidemiol 2021 Oct;74:101997 [FREE Full text] [doi: 10.1016/j.canep.2021.101997] [Medline: 34385076]
21. Figar S, Waisman G, De Quiros FG, Galarza C, Marchetti M, Loria GR, et al. Narrowing the gap in hypertension: effectiveness of a complex antihypertensive program in the elderly. Dis Manag 2004;7(3):235-243. [doi: 10.1089/dis.2004.7.235] [Medline: 15669583]
22. Lakkis NA, Atfeh AM, El-Zein YR, Mahmassani DM, Hamadeh GN. The effect of two types of SMS-texts on the uptake of screening mammogram: a randomized controlled trial. Prev Med 2011 Oct;53(4-5):325-327. [doi: 10.1016/j.ypmed.2011.08.013] [Medline: 21871480]
23. Were MC, Nyandiko WM, Huang KT, Slaven JE, Shen C, Tierney WM, et al. Computer-generated reminders and quality of pediatric HIV care in a resource-limited setting. Pediatrics 2013 Mar;131(3):e789-e796 [FREE Full text] [doi: 10.1542/peds.2012-2072] [Medline: 23439898]
24. Li Y, Xie YM, Huo J, Zhang H. [Data analysis of electronic medical recored clinical changes in indexs of liver and kidney function when salvianolate injection is parenterally administered over extended period]. Zhongguo Zhong Yao Za Zhi 2014 Sep;39(18):3593-3598. [Medline: 25532402]
25. Zhang HX, Li X, Huo HQ, Liang P, Zhang JP, Ge WH. Pharmacist interventions for prophylactic antibiotic use in urological inpatients undergoing clean or clean-contaminated operations in a Chinese hospital. PLoS One 2014 Feb 25;9(2):e88971 [FREE Full text] [doi: 10.1371/journal.pone.0088971] [Medline: 24586465]
26. Ghadieh AS, Hamadeh GN, Mahmassani DM, Lakkis NA. The effect of various types of patients' reminders on the uptake of pneumococcal vaccine in adults: a randomized controlled trial. Vaccine 2015 Oct 26;33(43):5868-5872. [doi: 10.1016/j.vaccine.2015.07.050] [Medline: 26232345]

27. Oluoch T, Katana A, Kwaro D, Santas X, Langat P, Mwalili S, et al. Effect of a clinical decision support system on early action on immunological treatment failure in patients with HIV in Kenya: a cluster randomised controlled trial. Lancet HIV 2016 Feb;3(2):e76-e84 [FREE Full text] [doi: 10.1016/S2352-3018(15)00242-8] [Medline: 26847229]
28. Wang Z, Hao G, Wang X, Wang W, Chen W, Zhu M. Short-term hypertension management in community is associated with long-term risk of stroke and total death in China: a community controlled trial. Medicine (Baltimore) 2016 Nov;95(48):e5245 [FREE Full text] [doi: 10.1097/MD.0000000000005245] [Medline: 27902588]
29. Al-Hashar A, Al-Zakwani I, Eriksson T, Sarakbi A, Al-Zadjali B, Al Mubaihsi S, et al. Impact of medication reconciliation and review and counselling, on adverse drug events and healthcare resource use. Int J Clin Pharm 2018 Oct;40(5):1154-1164. [doi: 10.1007/s11096-018-0650-8] [Medline: 29754251]
30. George G, Chetty T, Strauss M, Inoti S, Kinyanjui S, Mwai E, et al. Costing analysis of an SMS-based intervention to promote HIV self-testing amongst truckers and sex workers in Kenya. PLoS One 2018 Jul 6;13(7):e0197305 [FREE Full text] [doi: 10.1371/journal.pone.0197305] [Medline: 29979704]
31. Mody A, Sikazwe I, Czaicki NL, Wa Mwanza M, Savory T, Sikombe K, et al. Estimating the real-world effects of expanding antiretroviral treatment eligibility: evidence from a regression discontinuity analysis in Zambia. PLoS Med 2018 Jun 5;15(6):e1002574 [FREE Full text] [doi: 10.1371/journal.pmed.1002574] [Medline: 29870531]
32. Bachmann MO, Bateman ED, Stelmach R, Cruz AA, Pacheco de Andrade M, Zonta R, et al. Effects of PACK guide training on the management of asthma and chronic obstructive pulmonary disease by primary care clinicians: a pragmatic cluster randomised controlled trial in Florianópolis, Brazil. BMJ Glob Health 2019 Dec 16;4(6):e001921 [FREE Full text] [doi: 10.1136/bmjgh-2019-001921] [Medline: 31908865]
33. Engelbrecht R, Plastow N, Botha U, Niehaus D, Koen L. The effect of an occupational therapy mental health day treatment centre on the use of inpatient services in the Western Cape, South Africa. Disabil Rehabil 2019 Aug;41(16):1974-1980. [doi: 10.1080/09638288.2018.1453873] [Medline: 29701509]
34. Ismail S, Al-Subhi A, Youssif E, Ahmed M, Almalki A, Seger DL, et al. Patient-centered pharmacist care in the hemodialysis unit: a quasi-experimental interrupted time series study. BMC Nephrol 2019 Nov 13;20(1):408 [FREE Full text] [doi: 10.1186/s12882-019-1577-6] [Medline: 31722680]
35. Kelvin EA, George G, Mwai E, Kinyanjui S, Romo ML, Odhiambo JO, et al. A randomized controlled trial to increase HIV testing demand among female sex workers in Kenya through announcing the availability of HIV self-testing via text message. AIDS Behav 2019 Jan;23(1):116-125 [FREE Full text] [doi: 10.1007/s10461-018-2248-5] [Medline: 30109456]
36. Lima MF, Mondadori LA, Chibana AY, Gilio DB, Giroud Joaquim EH, Michard F. Outcome impact of hemodynamic and depth of anesthesia monitoring during major cancer surgery: a before-after study. J Clin Monit Comput 2019 Jun;33(3):365-371. [doi: 10.1007/s10877-018-0190-8] [Medline: 30074124]
37. Phillips TK, Mogoba P, Brittain K, Gomba Y, Zerbe A, Myer L, et al. Long-term outcomes of HIV-infected women receiving antiretroviral therapy after transferring out of an integrated maternal and child health service in South Africa. J Acquir Immune Defic Syndr 2020 Mar 01;83(3):202-209 [FREE Full text] [doi: 10.1097/QAI.0000000000002236] [Medline: 31725060]
38. Tay KH, Ariffin F, Sim BL, Chin SY, Sobry AC. Multi-faceted intervention to improve the antibiotic prescriptions among doctors for acute URI and acute diarrhoea cases: the green zone antibiotic project. Malays J Med Sci 2019 Jul;26(4):101-109 [FREE Full text] [doi: 10.21315/mjms2019.26.4.12] [Medline: 31496899]
39. Wu X, Tian M, Zhang J, Yang M, Gong X, Liu Y, et al. The effect of a multidisciplinary co-management program for the older hip fracture patients in Beijing: a "pre- and post-" retrospective study. Arch Osteoporos 2019 Mar 22;14(1):43. [doi: 10.1007/s11657-019-0594-1] [Medline: 30903390]
40. Ali M, Chwastiak L, Poongothai S, Emmert-Fees KM, Patel SA, Anjana RM, et al. Effect of a collaborative care model on depressive symptoms and glycated hemoglobin, blood pressure, and serum cholesterol among patients with depression and diabetes in India: the INDEPENDENT randomized clinical trial. JAMA 2020 Aug 18;324(7):651-662 [FREE Full text] [doi: 10.1001/jama.2020.11747] [Medline: 32809002]
41. Yang G, Gao C, Cai J. Prevention of nasal ala pressure injuries with use of hydroactive dressings in patients with nasotracheal intubation of orthognathic surgery: a randomized controlled trial. J Wound Ostomy Continence Nurs 2020;47(5):484-488. [doi: 10.1097/WON.0000000000000675] [Medline: 32649485]
42. Puttkammer N, Simoni JM, Sandifer T, Chéry JM, Dervis W, Balan JG, et al. An EMR-based alert with brief provider-Led ART adherence counseling: promising results of the InfoPlus adherence pilot study among Haitian adults with HIV initiating ART. AIDS Behav 2020 Dec;24(12):3320-3336 [FREE Full text] [doi: 10.1007/s10461-020-02945-8] [Medline: 32715409]
43. Roy M, Bolton-Moore C, Sikazwe I, Mukumbwa-Mwenechanya M, Efronson E, Mwamba C, et al. Participation in adherence clubs and on-time drug pickup among HIV-infected adults in Zambia: a matched-pair cluster randomized trial. PLoS Med 2020 Jul 1;17(7):e1003116 [FREE Full text] [doi: 10.1371/journal.pmed.1003116] [Medline: 32609756]
44. Kalichman S, Banas E, Kalichman M, Dewing S, Jennings K, Daniels J, et al. Brief enhanced partner notification and risk reduction counseling to prevent sexually transmitted infections, Cape Town, South Africa. Sex Transm Dis 2021 Mar 01;48(3):174-182 [FREE Full text] [doi: 10.1097/OLQ.0000000000001295] [Medline: 32976362]
45. Mody A, Sikazwe I, Namwase AS, Wa Mwanza M, Savory T, Mwila A, et al. Effects of implementing universal and rapid HIV treatment on initiation of antiretroviral therapy and retention in care in Zambia: a natural experiment using regression

discontinuity. Lancet HIV 2021 Dec;8(12):e755-e765 [FREE Full text] [doi: 10.1016/S2352-3018(21)00186-7] [Medline: 34656208]
46. Xu S, Wang X, Song Z, Han F, Zhang C. Impact and barriers of a pharmacist-led practice with computerized reminders on intravenous to oral antibiotic conversion for community-acquired pneumonia inpatients. J Clin Pharm Ther 2021 Aug;46(4):1055-1061. [doi: 10.1111/jcpt.13397] [Medline: 34101230]
47. Abdulrahman SA, Rampal L, Ibrahim F, Radhakrishnan AP, Kadir Shahar H, Othman N. Mobile phone reminders and peer counseling improve adherence and treatment outcomes of patients on ART in Malaysia: a randomized clinical trial. PLoS One 2017 May 16;12(5):e0177698. [doi: 10.1371/journal.pone.0177698] [Medline: 28520768]
48. Ye J, Orji IA, Baldridge AS, Ojo TM, Shedul G, Ugwuneji EN, et al. Characteristics and patterns of retention in hypertension care in primary care settings from the hypertension treatment in Nigeria program. JAMA Netw Open 2022 Sep 01;5(9):e2230025 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.30025] [Medline: 36066896]
49. Ye J, Li N, Lu Y, Cheng J, Xu Y. A portable urine analyzer based on colorimetric detection. Anal Methods 2017;9(16):2464-2471. [doi: 10.1039/C7AY00780A]
50. Zheng C, Yitong J, Zipu J, Tao W, Fang L. The long-term outcome of 3-dimensional CT-guided percutaneous radiofrequency thermocoagulation for tumor-related trigeminal neuralgia. Pain Physician 2019 Sep;22(5):E467-E475 [FREE Full text] [Medline: 31561659]
51. Ye J, Ren Z. Examining the impact of sex differences and the COVID-19 pandemic on health and health care: findings from a national cross-sectional study. JAMIA Open 2022 Sep 28;5(3):ooac076 [FREE Full text] [doi: 10.1093/jamiaopen/ooac076] [Medline: 36177395]
52. Ye J, Yao L, Shen J, Janarthanam R, Luo Y. Predicting mortality in critically ill patients with diabetes using machine learning and clinical notes. BMC Med Inform Decis Mak 2020 Dec 30;20(Suppl 11):295 [FREE Full text] [doi: 10.1186/s12911-020-01318-4] [Medline: 33380338]
53. Ye J. Pediatric mental and behavioral health in the period of quarantine and social distancing with COVID-19. JMIR Pediatr Parent 2020 Jul 28;3(2):e19867 [FREE Full text] [doi: 10.2196/19867] [Medline: 32634105]
54. Ye J. Advancing mental health and psychological support for health care workers using digital technologies and platforms. JMIR Form Res 2021 Jun 30;5(6):e22075 [FREE Full text] [doi: 10.2196/22075] [Medline: 34106874]
55. Ye J, Hai J, Wang Z, Wei C, Song J. Leveraging natural language processing and geospatial time series model to analyze COVID-19 vaccination sentiment dynamics on Tweets. JAMIA Open 2023 Apr 12;6(2):ooad023 [FREE Full text] [doi: 10.1093/jamiaopen/ooad023] [Medline: 37063408]
56. Ye J, He L, Beestrum M. Implications for implementation and adoption of telehealth in developing countries: a systematic review of China's practices and experiences. NPJ Digit Med 2023 Sep 18;6(1):174 [FREE Full text] [doi: 10.1038/s41746-023-00908-6] [Medline: 37723237]
57. Ferry A, Davis MJ, Rumprecht E, Nigro AL, Desai P, Hollier LH. Medical documentation in low- and middle-income countries: lessons learned from implementing specialized charting software. Plast Reconstr Surg Glob Open 2021 Jun 22;9(6):e3651 [FREE Full text] [doi: 10.1097/GOX.0000000000003651] [Medline: 34168942]
58. Syzdykova A, Malta A, Zolfo M, Diro E, Oliveira JL. Open-source electronic health record systems for low-resource settings: systematic review. JMIR Med Inform 2017 Nov 13;5(4):e44 [FREE Full text] [doi: 10.2196/medinform.8131] [Medline: 29133283]
59. Ye J. Patient safety of perioperative medication through the lens of digital health and artificial intelligence. JMIR Perioper Med 2023 May 31;6:e34453 [FREE Full text] [doi: 10.2196/34453] [Medline: 37256663]
60. Jalali R, Nogueira-Rodrigues A, Das A, Sirohi B, Panda PK. Drug development in low- and middle-income countries: opportunity or exploitation? Am Soc Clin Oncol Educ Book 2022 Apr;42(42):1-8 [FREE Full text] [doi: 10.1200/EDBK_10033] [Medline: 35658520]
61. Sanuade OA, Ale BM, Baldridge AS, Orji IA, Shedul GL, Ojo TM, et al. Fixed-dose combination therapy-based protocol compared with free pill combination protocol: results of a cluster randomized trial. J Clin Hypertens (Greenwich) 2023 Feb;25(2):127-136 [FREE Full text] [doi: 10.1111/jch.14632] [Medline: 36660886]
62. Ye J, Woods D, Bannon J, Bilaver L, Kricke G, McHugh M, et al. Identifying contextual factors and strategies for practice facilitation in primary care quality improvement using an informatics-driven model: framework development and mixed methods case study. JMIR Hum Factors 2022 Jun 24;9(2):e32174 [FREE Full text] [doi: 10.2196/32174] [Medline: 35749211]
63. Cowie MR, Blomster JI, Curtis LH, Duclaux S, Ford I, Fritz F, et al. Electronic health records to facilitate clinical research. Clin Res Cardiol 2017 Jan;106(1):1-9 [FREE Full text] [doi: 10.1007/s00392-016-1025-6] [Medline: 27557678]
64. Verheij RA, Curcin V, Delaney BC, McGilchrist MM. Possible sources of bias in primary care electronic health record data use and reuse. J Med Internet Res 2018 May 29;20(5):e185 [FREE Full text] [doi: 10.2196/jmir.9134] [Medline: 29844010]
65. Ye J. The impact of electronic health record-integrated patient-generated health data on clinician burnout. J Am Med Inform Assoc 2021 Apr 23;28(5):1051-1056 [FREE Full text] [doi: 10.1093/jamia/ocab017] [Medline: 33822095]

66. Ye J, Zhang R, Bannon JE, Wang AA, Walunas TL, Kho AN, et al. Identifying practice facilitation delays and barriers in primary care quality improvement. J Am Board Fam Med 2020 Sep 28;33(5):655-664 [FREE Full text] [doi: 10.3122/jabfm.2020.05.200058] [Medline: 32989060]
67. Mentz RJ, Hernandez AF, Berdan LG, Rorick T, O'Brien EC, Ibarra JC, et al. Good clinical practice guidance and pragmatic clinical trials: balancing the best of both worlds. Circulation 2016 Mar 01;133(9):872-880 [FREE Full text] [doi: 10.1161/CIRCULATIONAHA.115.019902] [Medline: 26927005]
68. Ye J. Design and development of an informatics-driven implementation research framework for primary care studies. AMIA Annu Symp Proc 2021;2021:1208-1214 [FREE Full text] [Medline: 35308925]
69. Global strategy on digital health 2020-2025. World Health Organization. 2021. URL: https://www.who.int/docs/default-source/documents/gs4dhdaa2a9f352b0445bafbc79ca799dce4d.pdf [accessed 2023-10-25]
70. Ye J. Health information system's responses to COVID-19 pandemic in China: a national cross-sectional study. Appl Clin Inform 2021 Mar;12(2):399-406 [FREE Full text] [doi: 10.1055/s-0041-1728770] [Medline: 34010976]
71. Join Us register. The George Institute for Global Health. URL: https://www.georgeinstitute.org/join-us-register [accessed 2023-10-16]
72. Mc Cord KA, Ewald H, Ladanie A, Briel M, Speich B, Bucher HC, et al. Current use and costs of electronic health records for clinical trial research: a descriptive study. CMAJ Open 2019 Feb 03;7(1):E23-E32 [FREE Full text] [doi: 10.9778/cmajo.20180096] [Medline: 30718353]

Abbreviations

EHR: electronic health record
HIC: high-income country
LMIC: low- and middle-income countries
PICOS: patients, problem, or population; issue of interest or intervention; comparison, control, or comparator; outcome; and study type
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
RCT: randomized controlled trial

Edited by C Lovis; submitted 06.03.23; peer-reviewed by D Tran, C Gao; comments to author 28.07.23; revised version received 10.09.23; accepted 22.09.23; published 22.11.23.

Please cite as:
Ye J, Xiong S, Wang T, Li J, Cheng N, Tian M, Yang Y
The Roles of Electronic Health Records for Clinical Trials in Low- and Middle-Income Countries: Scoping Review
JMIR Med Inform 2023;11:e47052
URL: https://medinform.jmir.org/2023/1/e47052
doi:10.2196/47052
PMID:37991820

Review

Patient Information Summarization in Clinical Settings: Scoping Review

Daniel Keszthelyi[1,2], MSc; Christophe Gaudet-Blavignac[1,2], PhD; Mina Bjelogrlic[1,2], PhD; Christian Lovis[1,2], MPH, MD

[1]Division of Medical Information Sciences, University Hospitals of Geneva, Geneva, Switzerland

[2]Department of Radiology and Medical Informatics, University of Geneva, Geneva, Switzerland

Corresponding Author:
Daniel Keszthelyi, MSc
Division of Medical Information Sciences, University Hospitals of Geneva
Rue Gabrielle-Perret-Gentil 4
Geneva, 1205
Switzerland
Phone: 41 223726201
Email: Daniel.Keszthelyi@unige.ch

Abstract

Background: Information overflow, a common problem in the present clinical environment, can be mitigated by summarizing clinical data. Although there are several solutions for clinical summarization, there is a lack of a complete overview of the research relevant to this field.

Objective: This study aims to identify state-of-the-art solutions for clinical summarization, to analyze their capabilities, and to identify their properties.

Methods: A scoping review of articles published between 2005 and 2022 was conducted. With a clinical focus, PubMed and Web of Science were queried to find an initial set of reports, later extended by articles found through a chain of citations. The included reports were analyzed to answer the questions of where, what, and how medical information is summarized; whether summarization conserves temporality, uncertainty, and medical pertinence; and how the propositions are evaluated and deployed. To answer how information is summarized, methods were compared through a new framework "collect—synthesize—communicate" referring to information gathering from data, its synthesis, and communication to the end user.

Results: Overall, 128 articles were included, representing various medical fields. Exclusively structured data were used as input in 46.1% (59/128) of papers, text in 41.4% (53/128) of articles, and both in 10.2% (13/128) of papers. Using the proposed framework, 42.2% (54/128) of the records contributed to information collection, 27.3% (35/128) contributed to information synthesis, and 46.1% (59/128) presented solutions for summary communication. Numerous summarization approaches have been presented, including extractive (n=13) and abstractive summarization (n=19); topic modeling (n=5); summary specification (n=11); concept and relation extraction (n=30); visual design considerations (n=59); and complete pipelines (n=7) using information extraction, synthesis, and communication. Graphical displays (n=53), short texts (n=41), static reports (n=7), and problem-oriented views (n=7) were the most common types in terms of summary communication. Although temporality and uncertainty information were usually not conserved in most studies (74/128, 57.8% and 113/128, 88.3%, respectively), some studies presented solutions to treat this information. Overall, 115 (89.8%) articles showed results of an evaluation, and methods included evaluations with human participants (median 15, IQR 24 participants): measurements in experiments with human participants (n=31), real situations (n=8), and usability studies (n=28). Methods without human involvement included intrinsic evaluation (n=24), performance on a proxy (n=10), or domain-specific tasks (n=11). Overall, 11 (8.6%) reports described a system deployed in clinical settings.

Conclusions: The scientific literature contains many propositions for summarizing patient information but reports very few comparisons of these proposals. This work proposes to compare these algorithms through how they conserve essential aspects of clinical information and through the "collect—synthesize—communicate" framework. We found that current propositions usually address these 3 steps only partially. Moreover, they conserve and use temporality, uncertainty, and pertinent medical aspects to varying extents, and solutions are often preliminary.

(JMIR Med Inform 2023;11:e44639) doi:10.2196/44639

KEYWORDS

summarization; electronic health records; EHR; medical record; visualization; dashboard; natural language processing

Introduction

Background

Summarization is an essential element of human cognition and consists of taking a set of information and retaining the pertinent elements to take action [1]. Feblowitz et al [2] defined information summarization in the health care context as "the act of collecting, distilling, and synthesizing patient information for the purpose of facilitating any of a wide range of clinical tasks." This definition translates as simplifying the presented information so that health care professionals (HCPs) can act more smoothly and efficiently in different clinical situations.

Automatic summarization of information in electronic health records (EHRs) can serve as a solution for information overload [3], a widespread problem in health care when the presented data are too much to be efficiently processed in a care situation. Information overload can have detrimental effects on patient care in the form of professional stress, fatigue, delays, and medical errors [4]. Although the phenomenon is not novel, it is increasingly present owing to an aging population with an exponentially increasing presence of chronic diseases, increased administrative burden, overabundance, and suboptimal storage of medical data [5,6]. Furthermore, current EHR systems present information in a fragmented manner [7] with widespread repetition, copy-pasting [8], and details not relevant to clinical care [9].

Despite the need for automatic patient information summarization, there is no widely accepted theory or methodology. This report aimed to synthesize the contributions of patient information summarization scattered in the literature. Scoping review is the chosen form with the aim of mapping ideas, mapping concepts related to the question, and identifying knowledge gaps.

The review is not unprecedented: in their narrative review, Pivovarov and Elhadad [10] already summarized the most important contributions to clinical summarization in 2015. Moreover, there have been several published studies surveying the literature in related fields, including the summarization of biomedical literature [11,12], the summarization from medical documents [13], neural natural language processing (NLP) in EHRs [14], named entity recognition, a type of information extraction and NLP technique, free-text clinical notes [15], automatic clinical documentation [16], the visualization of medical information in the clinical context [17-20], the visualization of intensive care unit (ICU) data [21], and the visualization of trends in medical data [22]. The latter reviews, although exhaustive in their specific scope, do not permit the identification of state-of-the-art summarization methods for HCPs. For example, it is difficult to state the current state of research for the management of uncertainty and time in clinical summarization. Moreover, they did not provide any informed guidelines for clinical summarization.

Objective

This review, building on a broader scope of articles than the combination of all the previous studies, systematically evaluates where, what, and how medical information is summarized; whether summarization conserves temporality, uncertainty, and medical pertinence; and how the propositions are evaluated and deployed.

On the basis of cognitive science literature, this review also proposes a novel "collect—synthesize—communicate" framework to compare studies on how they contribute to clinical summarization.

Methods

Overview

The methodology was designed to process a broad scope of articles; hence, different search strategies were combined to diversify the sources. Two reviewers agreed on the selection method: 2 databases with a clinical focus were searched with similar queries and the retrieved articles were filtered by one of the reviewers according to their titles and abstracts. The same reviewer read the remaining reports in the full text and selected them according to the inclusion and exclusion criteria. The same filtering was then carried out on citations within these articles and the citations of these articles. The reporting was done using the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) [23], and a checklist is provided in Multimedia Appendix 1.

The 2 databases searched in this review were the Web of Science Core Collection and PubMed as they contain a broad scope of articles in the medical field and are less inclusive of other articles in computer science, not related to the medical or scientific domain.

The search query for Web of Science was designed as a combination of 2 parts: capturing the summarization process and capturing the health care content. An iterative process was used to define the exact search term, where the gain of adding a keyword was examined by determining whether the first 5% (sorted by relevance defined by Web of Science) of the results from a query containing the new word but excluding the previous words shows any relevant article. This led to the following query: "ALL=((‘ehr’ OR ‘emr’ OR ‘health’ OR ‘patient’ OR ‘medical’ OR ‘hospital’ OR ‘healthcare’) AND (‘summarization’ OR ‘summarisation’ OR ‘summarizing’))" searching in the title, abstract, and metadata of the articles in the database, including the "keyword plus" field containing terms frequently appearing in the body of an article but not mentioned in the title or abstract.

The query in PubMed was "(‘her’ OR 'emr' OR ‘health’ OR ‘patient’ OR ‘medical’ OR ‘hospital’ OR ‘health care’ OR ‘medical record’[(MeSH Terms])) AND (‘summarization’ OR ‘summarisation’ OR ‘summarizing’)." This query was almost identical to the search made in Web of Science except that

PubMed does not have a "keyword plus" field like Web of Science to search in. Instead, preindexed articles with the Medical Subject Headings term "medical records" were included in the search.

All results of the queries were imported into the Rayyan app [24], helping to organize the citations for a review article. After duplicates were removed, the abstracts and titles were scanned in this application to filter records that were obviously irrelevant to patient information summarization. The app enables highlighting specific words in the abstract with 2 distinct colors, speeding up the review process. After this filtering step, the remaining articles were read in the full text for inclusion using the inclusion and exclusion criteria detailed in the following section.

After identifying the relevant works, the list of references in the selected articles and the list of citing papers retrieved by Google Scholar were manually reviewed for titles related to the topic. These potentially relevant titles were manually added to the citation manager and further filtered by reading their abstract and eventually reading them in the full text (as with the original results). If these articles contained further (previously unseen) relevant references or citations toward them, they were also processed.

Inclusion and Exclusion Criteria

According to the inclusion criteria, all records mentioning the summarization of clinical or health data as a general goal in their abstract, proposing solutions for information overflow, or claiming to make steps for these general goals were included.

All records were excluded where the corresponding article was not available in the full text for the authors (ie, at the University of Geneva campus), as the analysis could not be conducted on these records.

As several articles do not mention the source of the data in their abstracts, an exclusion criterion excludes articles about summarizing non-EHR data (eg, summarizing research articles; not EHR). Similarly, articles developing summarization for users other than HCPs *(not for HCP)* or for contexts other than clinical applications *(not clinical)* were excluded.

As many different methods can be labeled as "summarization," only records presenting a type of overview of a patient's current or past status are aimed to be included, therefore articles proposing alerts (eg, risk scores and cues) or similar simple parameters to summarize the state of a disease or patient *(alert)* and articles proposing other remedies for information overflow than automatic summarization for information overflow *(not automatic)* were excluded.

As previous reviews analyzed articles using different aspects, a broad timeframe was aimed at the review. However, as early EHR systems were very different from current systems, and hence the concept of summarization is largely different in these systems, articles before 2005 were excluded (*<2005*). The cutoff year is somewhat of an arbitrary (but round) threshold, although contributions before this year are sporadic.

Finally, articles presenting summarization solutions only for nontextual and nonstructured data (eg, video or signal summarization; *Other data)* and review papers *(Review)* were also excluded.

Articles found relevant to the review were evaluated by one of the reviewers for several criteria chosen to answer the following questions:

1. Where is summarization performed?
2. What is summarized? How?
3. How crucial aspects of clinical information are conserved and used?
4. How are the algorithms evaluated?

Textbox 1 presents the detailed criteria. Some of these criteria were defined a priori, whereas others were shaped during the analysis process. For 1 aspect, the input data type for summarization, the analysis was carried out on a broader scope, and reports excluded by the "other data" criterion were also analyzed for this information.

Textbox 1. Criteria according to which articles are evaluated in the analysis part of this review. For some of the criteria, categories are defined a priori. For others, they are shaped during the analysis process (the ones defined a posteriori are marked with an asterisk).

General aspects

- Type of the study
  - Prototype: articles describing a summarization system or algorithm that can be evaluated. The evaluation might be present or absent from the article.
  - Evaluation study: articles evaluating summarization systems, algorithms, or current summarization processes in health care without presenting a new automatic summarization solution.
  - Recommendations: articles with theoretical contributions not being implemented.

Where are summaries needed?

- Field of application*: the medical or clinical domain where the summarization is applied. The categories are discovered during the review process.

What should be summarized?

- Source of information:
  - Single encounter
  - Multiple encounters
  - This information cannot be inferred from the text.
- Input for the summarization:
  - Structural data: a combination of numerical and categorical data
  - Textual data: free-text patient information present in electronic health record systems

How to summarize?

- The summarization method*: the categories of summarization methods are shaped during the review process.
- Presentation*: how the summary is presented to the end user. The types of presentations are shaped during the review process.
- View on the summarization problem:
  - The top-down group represents records where summarization consists of eliminating “disturbances” from all available information, that is, hiding information deemed to be unnecessary.
  - Bottom-up methods see summarization as a process of finding the most salient information available and building a summary from it.

Aspects to be conserved during summarization

- Temporality*: if and how temporal information is conserved and used during summarization. The categories are shaped by the discoveries in the scoping review.
- Uncertainty*: if and how the uncertainty of information is represented during summarization. The categories are shaped by the discoveries in the scoping review.
- Medical knowledge*: if any medical knowledge is included in the design of the summarization system or during summarization. The categories are shaped during the review process.

What is a good summary? Evaluation and deployment

- Evaluation*: the method of evaluation. The types are shaped according to the discoveries from the review process.
- Deployment: if the summarization system was deployed in real clinical settings

Collect—Synthesize—Communicate Framework

During the analysis, we developed a new framework to compare methods of how they summarize clinical information. Following the definition presented in the introduction [2], the model divides the summarization process into an ideally sequential process of information collection, information synthesis, and summary communication. Information collection refers to the extraction of information from raw data, synthesis describes the selection and eventual transformation of the retrieved information, and communication refers to the representation of the synthesized information in a human-digestible format.

This view was consistent with that of several sources of cognitive psychology. For example, Johnson [25] describes summarization as a sequence of prerequisites for summarization (including comprehending individual propositions of a story,

establishing connections, identifying the consistent structure of the story, and remembering the information), information selection, and formulating a concise summary. This is also similar to the view presented by Hidi and Anderson [26], who discussed selection, condensation, and transformation.

Nevertheless, the few theoretical studies on clinical summarization have slightly different views on the process. Feblowitz et al [2] described clinical summarization as a process of aggregation, organization, reduction, transformation, interpretation, and synthesis. Jones [27,28] describes textual summarization as a process of interpretation, transformation, and text generation. Although these theories mention seemingly different steps for summary creation, they can be mapped to the proposed simpler and more general 3-step framework. Table 1 presents the mapping to the present framework of these theories and some of the most commonly used summarization methods.

Table 1. Summary of how existing theoretical frameworks and most abundant summarization methods relate to the collect—synthesize—communicate framework.

Theory or method	Collection	Synthesis	Communication
Feblowitz et al [2]	Aggregation, organization, and interpretation	Reduction, transformation, and synthesis	Organization and synthesis
Jones [27,28]	Interpretation	Transformation	Text generation
Extractive summarization (eg, Liang et al [29])	N/A[a] (not covered)	Sentence selection	N/A (not covered)
Abstractive summarization (eg, Gundogdu et al [30])	Encoding	Attention mechanism	N/A (not covered)
Topic modeling (eg, Botsis et al [5])	Topic extraction	N/A (not covered)	N/A (not covered)

[a]N/A: not applicable.

Results

Overview

As shown in Multimedia Appendix 2 [31], a total of 7925 titles were retrieved from PubMed and 3641 articles were retrieved from Web of Science. After removing duplicates, 9166 records were screened by their title and abstract for inclusion criteria and 380 records were chosen for full-text reading. From these, 1 could not be accessed by the authors and 328 were excluded based on the exclusion criteria.

From the 52 articles included in the analysis, 612 records were identified as potentially relevant by their title and 175 titles were chosen to be read in the full text after screening the abstracts. From these 175 titles, 2 could not be accessed, 97 records were excluded according to the exclusion criteria, and 76 titles were included in the analysis.

Among the 128 articles remaining in the analysis, 102 titles were categorized as a prototype, 20 were categorized as evaluation studies, and 6 were categorized as "recommendations."

Fields of Application

This review identified diverse fields of application for which summarization methods have been developed. A grouping of these fields is as follows:

1. ICU (27/128, 21.1%), where recent events and vital parameters are summarized
2. *Surgery* (1/128, 0.8%) and *related anesthesiology* (5/128, 3.9%), requiring all the information related to surgery to be summarized
3. *Diagnostics*, showing findings from one or several diagnostic sessions and including radiology (19/128, 14.8%), out of which 5.5% (7/128) were presented as a solution in the MEDIQA 2021 summarization task [32], ultrasound (2/128, 1.6%), prostatectomy (1/128, 0.8%), and laboratory data management in a clinical context (1/128, 0.8%)
4. Hospital care (9/128, 7%), where information related to a hospital stay requires efficient summarization
5. Chronic disease monitoring including diabetes (4/128, 3.1%), HIV (1/128, 0.8%), chronic obstructive pulmonary disease care (1/128, 0.8%), cardiology (2/128, 1.6%), nephrology (1/128, 0.8%), and monitoring of multiple chronic diseases (4/128, 3.1%), where salient events and information during a complex and long-lasting disease are required
6. Oncology (5/128, 3.9%), where the main events and elements of complex treatment are summarized
7. Drug prescription (3/128, 2.3%), where pharmaceutical history is summarized
8. Other medical environments included psychotherapy (3/128, 2.3%), opioid misuse treatment (1/128, 0.8%), general practice (2/128, 1.6%), emergency room (2/128, 1.6%), older adult care (2/128, 1.6%), and maternal care (1/128, 0.8%).

In addition, 25% (32/128) of articles did not specify their field of application or were meant to be usable in multiple types of medical environments and domains.

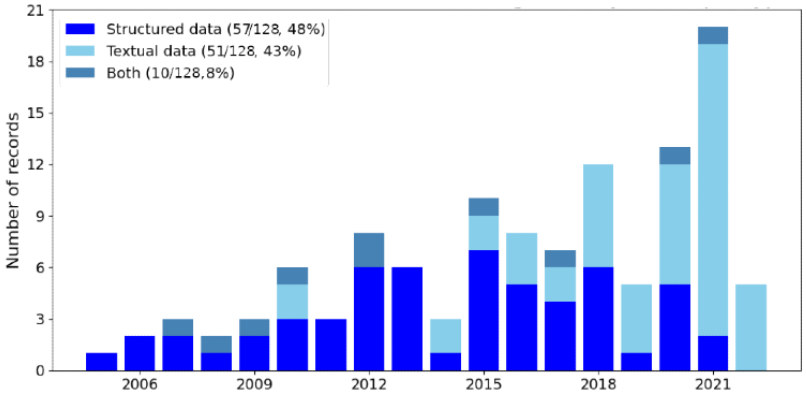
Input for Summarization

Regarding the source of information, 62.5% (80/128) of reports talk about systems summarizing single patient encounters, 27.3% (35/128) of reports explicitly talk about summarizing multiple encounters, 6.3% (8/128) of reports implicitly describe multiple encounter summarization, and 3.9% (5/128) of reports did not specify the cardinality of encounters.

Among the 128 articles in the review, 3 (2.3%) reports *do not specify the input* type for the summary, 59 (46.1%) worked only with *structured data*, 53 (41.4%) worked only with *textual data*, and 13 (10.2%) worked with *both types*. The trends in the number of articles with different input types are shown in Figure

1. Although more records use only structured data as the input type, the number of articles using textual information has shown a rapidly increasing trend in recent years. Textual information is usually assumed to be in English [33,34] and presents solutions for Finish and German languages [35,36].

Figure 1. The number of records by year of publication and the input type used in the summarization system or method presented or evaluated. Each column corresponds to a year, the different input types are aggregated into this column, their proportion for the given year is visible on the figure. ICU: intensive care unit.



A rapid analysis of the excluded articles using other types of clinical data identified the following:

1. Overall, 37 video and image sequence visualization tools in the clinical domain, mainly keyframe extraction [37,38] or motif discovery [39] methods in various fields of medicine, including older adult care, endoscopy [40], hysteroscopy [40,41], laparoscopy [42], magnetic resonance image [39,43], and ultrasound [44]
2. Overall, 20 sensory data simplification techniques using time-series analysis, motif discovery [45], and classification [46] methods for electrocardiogram or other types of signals
3. Overall, 117 articles about summarizing genomic data [47]

How Data Can Be Summarized? Summarization Methods

Common summarization methods used in the analyzed studies include (a report might use several of these) the following:

1. *Visual design* (59/128, 46.1%) organizes the information visually to help HCPs understand it within a short timeframe.
2. *Concept and relation extraction* (30/128, 23.4%): extracts semantic information from textual information
3. *Abstractive summarization* (19/128, 14.8%) [30,48-52] shortens texts by reformulating them using different wording to describe the content of a document [10].
4. *Extractive summarization* (13/128, 10.2%) [29,53] shortens texts by omitting a part of it, that is, composing a short text (a summary) from extracts of the original document.
5. *Summary specification* (11/128, 8.6%) describes the content to be presented for a summary.
6. *Pipeline extracting* information and *synthesizing and communicating* it with natural language generation tools (7/128, 5.5%)
7. *Topic modeling* (5/128, 3.9%) [54-57] categorizes documents according to their content and labels them with a list of representative words [58]
8. *Time-series analysis* (6/128, 4.7%) identifies characteristic properties in a temporal data series, including motif discovery, identifying meaningful patterns in temporal data (used in the study by Jane et al [59]), trend detection [60], or change detection [61].
9. *Dimensionality reduction* (3/128, 2.3%) treats patient data as a long vector encoding all patient information (ie, a row in a table with many columns), and reduces this information to a shorter vector (ie, a row with a much smaller number of columns) without losing too much information.

Some of these methods are intrinsic to the input data type and work only with a particular data type. For example, time-series analysis (including motif discovery), risk scores, and dimensionality reduction are intrinsic methods for structured data. Although a large number of articles using these methodologies are not included in this review as they are used by non-HCPs (eg, machine learning algorithms), some of the titles propose this approach as the first step to clinical summarization [62-64].

The most common intrinsic methods for textual data are extraction, abstractive summarization, and topic modeling.

Some of these summarization methods can apply machine learning techniques. An overview of the applied machine learning methods is presented in Table 2. The table lists all records obtained using machine learning and categorizes the records according to the summarization method and the type of machine learning method they use. Machine learning methods can be categorized into traditional machine learning methods, deep neural networks, and transformers. Traditional methods include support vector machines, random forests, and conditional random field methods; deep neural networks include deep neural networks, recurrent neural networks, and convolutional neural networks; transformers contain BART [65], BERT [66], Pegasus-based [67] methods, and pointer-generator models. In addition, Reunamo et al [34] used an interpretable machine learning technique (Local Interpretable Model-Agnostic Explanations, LIME [68]). N/A indicates that a machine learning method is not used for a given type of summarization method.

Table 2. Summary of records applying machine learning methods for clinical summarization[a].

	Traditional techniques	Deep neural networks	Transformers

Extractive summarization	• SVM[b]+CRF[c]
	  • Liang et al [29], 2019
	  • Liang et al [9], 2021
		• CNN[d]
		  • Liang et al [29], 2019
		  • Liang et al [9], 2021
		  • Subramanian et al [69], 2021
		• RNN[e]
		  • Alsentzer and Kim [70], 2018
		  • Liu et al [71], 2018
		  • Chen et al [53], 2019
			• BERT[f]
			  • Chen et al [53], 2019
			  • Kanwal and Rizzo [72], 2022
			  • McInerney et al [73], 2020
			  • Liang et al [36], 2022
			  • Shah and Mohammed [56], 2020
			• Other
			  • Liang et al [29], 2019
			  • Liang et al [9], 2021

Abstractive summarization	• N/A[g]	• RNN
		  • Gundogdu et al [30], 2021
		  • Hu et al [74,75], 2021
			• BERT
			  • Cai et al [48], 2021
			  • Chang et al [76], 2021
			  • Mahajan et al [77], 2021
			  • Sotudeh et al [50], 2020
			• BART[h]
			  • Dai et al [78], 2021
			  • He et al [79], 2021
			  • Kondadadi et al [80], 2021
			  • Shing et al [81], 2021
			  • Xu et al [82], 2021
			• Pegasus
			  • Dai et al [78], 2021
			  • He et al [79], 2021
			  • Kondadadi et al [80], 2021
			  • Zhu et al [89], 2021
			  • Xu et al [82], 2021
			• Pointer generator
			  • MacAvaney et al [49], 2019
			  • Zhang et al [51], 2018
			  • Zhang et al [83], 2019
			• Own architecture
			  • Delbrouck et al [84], 2021
			• GPT-2[i]
			  • Xu et al [85], 2019

Concept and relation extraction	• N/A	• RNN:
		  • Reunamo et al [34], 2022
			• BART:
			  • Tang et al [86], 2022

Pipeline	• Random forest: • Lee and Uppal [87], 2020	• N/A	• N/A
Topic modeling	• Alternating decision tree: • Devarakonda et al [88], 2017	• N/A	• N/A

[a]Machine learning methods are categorized into traditional machine learning methods, deep neural networks, and transformers.

[b]SVM: support vector machine.

[c]CRF: conditional random field.

[d]CNN: convolutional neural network.

[e]RNN: recurrent neural network.

[f]BERT: Bidirectional Encoder Representation from Transformers.

[g]N/A: not applicable.

[h]BART: Bidirectional Autoregressive Transformer.

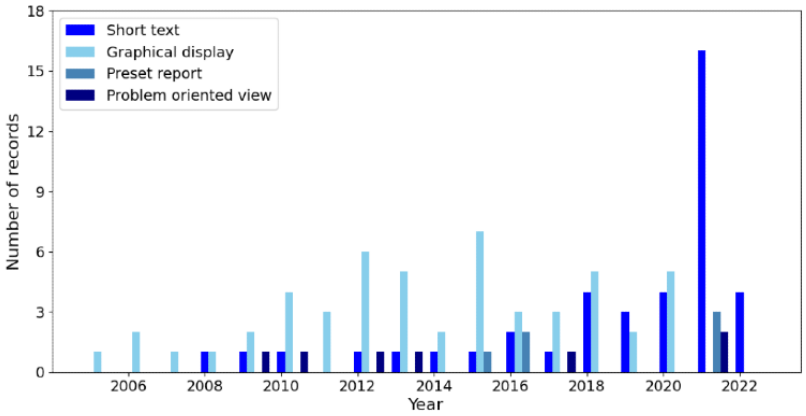
[i]GPT-2: Generative Pre-trained Transformer 2.

Summarization methods can also be categorized based on their outputs. The review identified several ways in which the summarized information is presented to the end user:

- A graphical display (53/128, 41.4%) is a specific way (interactive or not) to present information on the computer screen.
- A short textual summary (41/128, 32%) describes information in an ordinary language (eg, English).
- A preset static report: including its content designed to include specific medical information (6/128, 4.7%) or chosen statistical distributions representative of the patient (1/128, 0.8%)
- Problem-oriented view: a view grouping findings according to the problems the patient may present (7/128, 5.5%).
- Low-dimensional vector (4/128, 3.1%): encoding information n numbers (where n is the dimension) where each number represents the state of the patient from a particular aspect.
- List of words representing a topic (5/128, 3.9%), problem list (2/128, 1.6%), list of medical concepts found in the document (2/128, 1.6%), or label (2/128, 1.6%)
- A table (1/128, 0.8%) with rows and columns or a directed graph or concept map representing information in a graph-structured data model (2/128, 1.6%)
- No presentation: the articles in the "recommendation" group (5/128, 3.9%) did not present the results to the end user.

Figure 2 depicts the evolution (by the time of publication of records) of the most abundant formats for communicating the summarization results.

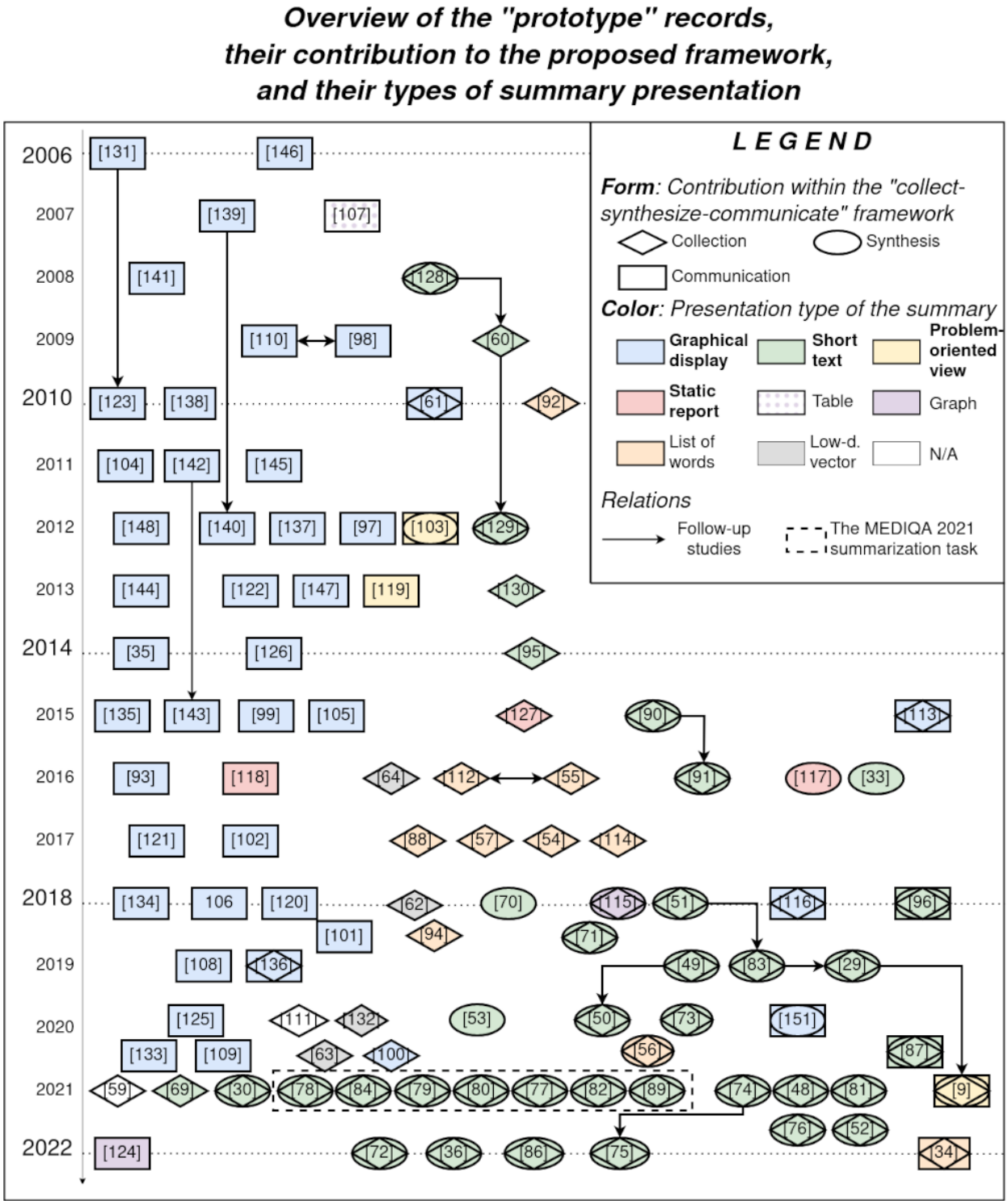
Figure 2. The number of records by year of publication and the most common ways of summary presentation used in the summarization method presented or evaluated in the report.



Concerning their view on summarization, 32.8% (42/128) of the records regarded summarization as a *bottom-up* approach and 64.8% (83/128) used the *top-down* view, whereas 1.6% (2/128) of records *do not show a clear opinion* on summarization.

Using the proposed framework, 42.2% (54/128) of the records contributed to *information collection*, 33.6% (43/128) recorded *information synthesis*, and 46.1% (59/128) presented solutions for *summary communication*. Figure 3 [9,29,30,33-36,48-57,59-64,69-84,86-149] visualizes all the analyzed prototype articles and how they fall into these categories (ie, which step of the framework is addressed within the corresponding work). The records' year of publication, presentation of summaries, and relationship between records are also displayed. The diagram has a vertical axis showing the year of publication, and all the "prototype" records (presented as a reference) published in that year appear in a line (or in 2 lines if the number of publications for a given year is very high). The order within a line has no significance, although the records were grouped within a line to show their contributions. The shape or shapes surrounding a reference symbolizes the steps that the record addresses in the "collect—synthesize—-communicate" framework. The reference to the study by Liang et al [9] is surrounded by all 3 shapes, indicating that the study addresses all the 3 steps. The records also have a color representing in which format the summaries are presented to the HCPs. Closer relationships (ie, follow-up studies) are also presented. The studies submitted to the MEDIQA-2021 challenge [21] are also marked in the diagram.

Figure 3. Diagram showing references to all analyzed "prototype" records and how they contribute to the "collect—synthesize—communicate" framework (ie, which step of the framework is addressed within the corresponding work). The records' year of publication, the presentation of summaries, and the relation between records are also displayed [9,29,30,33-36,48-57,59-64,69-84,86-149].



Regarding information collection, concept and relation extraction (50/54, 56%), time-series analysis (6/54, 11%), encoding (22/54, 41%), temporal abstraction (6/54, 11%), and topic extraction (5/54, 9%) were proposed as solutions. Medical concepts are extracted from textual data either using publicly available solutions (eg, cTAKES [164] in the study by Goff and Loehfelm [94]) or tools developed by the authors (eg, [113,114,157]). The retrieved list of concepts can be used for simpler tasks, such as problem list generation [88], or some records present systems that take a step further extracting the context [115], syntactic structure [94], or approximate semantic structure of a sentence [116] as well.

With regard to information synthesis, sentence selection by scoring (13/43, 30%), knowledge-based rules (18/43, 42%), and attention mechanism (19/43, 44%) were possible solutions.

Proposals for summary visualizations are usually features on a graphical screen; they are listed and compared in Table 3. For

unprocessed textual data, the solutions included highlighting important concepts (3/5, 60%) and creating graphs that visualize the semantic structure of the textual data (2/5, 40%).

Table 3. The number of records presenting various features for visualizations in works with graphical displays. A record can use several features.

Feature	Occurrence (out of 58), n
Colors	43
Selection of features	37
Tabular interface	35
Change in time	31
Visualization of divergence	22
Placement of variables	19
Interactive display	18
Pictograms	10
Physical location shown	6
Alerts	5
Size difference	5
Customizability	3
Shape	3
Word cloud	3
Comparison	2
Variability of parameters	1

Aspects to Be Conserved and Used

In total, 58.6% (75/128) of the titles did not conserve temporal information, whereas 2.3% (3/128) of titles were agnostic to temporal information (they conserve but do not use it). The remaining articles used a variety of approaches:

- *Timeline visualization*: plotting information along a horizontal or vertical temporal axis (34/128, 26.6%).
- *Other visualizations*: showing only the trend of parameters (1/128, 0.8%) or providing a complex visualization framework in which temporal information can be displayed and analyzed (1/128, 0.8%).
- *Information extraction from the temporal domain* by analyzing how the parameters change during the patient journey. This group included a time-series analysis (6/128, 4.7%), pattern recognition (2/128, 1.6%), and change detection (1/128, 0.8%). Time-series analysis (applied in several studies) [59,60,125-127,150] extracts statistical information from the temporal evolution of one or several variables. Pattern recognition [117,128] attempts to identify meaningful patterns in temporal data. Change detection [61] seeks important events that manifest in the trends and patterns of temporal variables. Some studies have revealed the relationship between these events.
- A *theoretical model of temporal events*, which can either describe more complex interactions between temporal events (6/128, 4.7%) or be very simple (eg, creating an order: 1/128, 0.8% or describing events with a single time [n=1]).

It is worth noting that timeline visualization was applied in 3 articles in the temporal information extraction and in 1 article in the complex model of the temporality group as well.

Regarding information uncertainty, 89.1% (114/128) of the articles did not consider the uncertainty of information. Others have proposed the following solutions:

- Statistical methods were used to treat *uncertainty in data*. These methods included correcting detectable errors (3/128, 2.3% [60,79,150]) and optimizing the statistical description of the data using robust statistics (1/128, 0.8% [62]).
- *Uncertainty of temporal events*was described (2/128, 1.6%).
- *Uncertainty of statements* was described by assigning them to uncertainty categories (3/128, 2.3% [71,74,75]) or using existing ontology (3/128, 2.3%).

Medical pertinence was not conserved in 34.4%, (44/128) of the studies (ie, they had no requirements that the summary had any relation to medical concepts or knowledge). A total of 35.9% (46/128) of records used medical knowledge to specify the information to be included in the summary and with what design. Other propositions included the following:

- Using ontologies to find and relate concepts within textual notes (20/128, 15.6% [87,92,93]), the use of Unified Medical Language System (UMLS) extraction tools (6/128, 4.7%) to extract them (eg, [94-96]), or improving the performance of abstractive summarization (2/128, 1.6% [76])
- Use of risk scores to create visualizations (3/128, 2.3% [97-99]) or the application of guidelines to assess risks (2/128, 1.6% [100,101])

- The use of medically salient rules to constrain summarization (3/128, 2.3% [29,61,102])
- Evaluation of the factual correctness of the created summaries integrated into reinforcement learning (2/128, 1.6% [79,83])
- Application of medical knowledge to select pertinent information (2/128, 1.6% [103,151])
- The use of medical knowledge to construct evaluation metrics (2/128, 1.6% [81,152])

What Is a Good Summary? Evaluation and Deployment

Several types of evaluation methods and metrics are presented in the publications:

- Quantitative measurements in experiments with human participants (31/128, 24.2%)
- Quantitative measurements when summarization was performed in a real clinical environment (8/128, 6.3%)
- Interviews (7/128, 5.5%), focus groups (2/128, 1.6%), or surveys (19/128, 14.8%) asking the opinions of the users after exposure to the summarization system
- Intrinsic evaluation (25/128, 19.5%) of measuring quality by comparing the results to a ground truth
- Performance on a proxy task (ie, disease prediction; 10/128, 7.8%)
- Performance in identifying human-annotated concepts (9/128, 7%) or topics (2/128, 1.6%)

The distribution of the number of human evaluators is shown in Multimedia Appendix 3. Two records [119,173] used significantly more evaluators than other solutions, which are represented as 2 distinct groups at the tail of the histogram. One of these records [173] is a large-scale survey, whereas the other [119] is a pilot study measuring user performance.

Although some records present several evaluation techniques, in 9.4% (12/128) of the articles, no evaluation is presented; in 1.6% (2/128) of articles, the evaluation is not detailed; and in 4.7% (6/128) of records, the evaluation consists of a subjective evaluation carried out by the authors of the article.

The metrics used in the evaluations are as follows:

- *Performance metrics* (eg, precision, recall, and *F* score) on a *prediction/classification task* measuring the "goodness" (validity) of predictions or classifications (used both in usability experiments and formative evaluations; 11/128, 8.6%)
- *Performance metrics* (eg, accuracy) of human participants (ie, the validity of their decisions) on an *experimental task* (24/128, 18.8%)
- *Time savings* due to summarization systems: *time to completion* (*ie, the time needed to perform a predefined task*) in experiments (21/128, 16.4%) or time saved during patient visits (1/128, 0.8%) in deployed systems
- *Patient outcome metrics* (6/128, 4.7%) included mortality and hospital readmission rates.
- The *NASA-TLX score* describes the workload of the user (5/128, 3.9%) and the relationship between the NASA-TLX score and error count (1/128, 0.8%).
- *Number of interactions* (eg, click and screen change) in usability studies (3/128, 2.3%)
- *Grades* given by human evaluators measuring the utility and usability of a system (13/128, 10.2%) or trust in it (1/128, 0.8%).
- Scores comparing textual summaries with properties of the input text. These scores included *Recall-Orientated Understudy for Gisting Evaluation (ROUGE)* [153] (20/128, 15.6%), *bilingual evaluation understudy* [154] (2/128, 1.6%), and *comparison between* input and output *distributions*(2/128, 1.6%).
- *Other properties of the output textual summaries* including *readability/fluency* (10/128, 7.8%), *accuracy* or *factual correctness* (5/128, 3.9%), *completeness* (7/128, 5.5%), and *overall quality* (7/128, 5.5%) in qualitative evaluations of textual outputs. Two (N=128, 1.6%) records distinguished between ontological and nonontological correctness.
- *Proxy measures for the faithfulness* of textual summarization (6/128, 4.7%)
- *Heuristics* derived from requirement specification (5/128, 3.9%)

The evaluation metrics used in quantitative evaluations usually depend on the method of summarization; for dimensionality reduction, it is often a performance metric to predict diseases; for extractive and abstractive summarizations, the ROUGE score [153] is the most commonly used metric, as it is considered the most reliable [32], and for topic modeling, it is its empirical likelihood [174].

For text summarization, evaluations with human participants are often carried out by annotators subjectively grading each produced summary along some metrics, including readability, factual correctness, and completeness. For other summarization methods, this task is usually approximated by either usability tests [134-139] or experiments [140-147] where performance and workload are measured. The few systems deployed in clinical settings are often evaluated by measuring patient outcomes or clinical indicators.

Reviewing the results of each report, some records compared the results with summarization methods in the general domain [48], and 6 (5%) [30,32,50,75,78,144] presented a comparison of clinical summarization methods. The distribution of cross-citations between articles, that is, the number of other publications appearing in the review cited by each report, is represented in Multimedia Appendix 4. Furthermore, 80% of the records cited fewer than 3 other articles analyzed in this review.

Among the 128 records analyzed, 4 (3.1%) talked about a method deployed on a large scale, 7 (5.5%) described deployment in a pilot study, and 1 (0.7%) disclosed the code alongside the publication.

Discussion

Principal Findings

Where Are Summaries Needed in Health Care?

Publications on clinical summarization are tied to several different medical and clinical fields, mainly where quick decision-making is crucial (eg, ICU) or where a large amount of information is routinely produced (eg, oncology, chronic disease management, and hospital care).

However, some fields requiring quick decision-making (eg, emergency room environments) have seen less progress. In contrast, others where quick decision-making is less critical (eg, radiology) are covered by a relatively large number of records. This discrepancy suggests that clinical summarization can be beneficial in almost all medical fields, although the idea may not have reached all domains at the same pace. Although the previous drivers are easily identifiable, we speculate that the presence of other solutions proposed to handle information overload (eg, the study by Xu et al [85]; see the study by Hall and Walton [4] for review) can decelerate, whereas a shortage of personnel in a field (eg, radiology [155]) can accelerate adaptation.

What Should Be Summarized?

The increasing trend in both single-encounter and multiencounter summarizations suggests that both types are salient and should be used depending on the care situation.

Regarding the input for summarization, several experiments show that HCPs can act at least as accurately and in a timely manner with summarized structured [104-110,156] data or textual data [60,157] or with most information coded in these types of data [158] than using complete documentation. Therefore, the focus should be on summarizing textual and structured data when creating summaries for HCPs.

The increasing trend of using textual data for summarization might be attributed to the improvement of NLP, the improved computing power required for some NLP tasks, and the results published by Van Vleck et al [158], who claimed that a significant portion of patient information lies in clinical notes. In contrast, Hsu et al [111] challenged this hypothesis by presenting experiments to predict some clinical measures (eg, hospital readmission and mortality) using textual and structured patient information sources. They concluded that textual sources have little predictive power for the outcomes. However, their analysis might be biased by their methodology, as they use only simple syntactic metrics to describe textual information, whereas semantic information is not included in their model.

How Data Can Be Summarized?

The records analyzed in this review show myriad techniques for summarizing clinical data. Some are intrinsic to the input data type and work only with a particular data type, whereas others are not dependent on the input data type.

For textual data, the review reveals more works about abstractive summarization than extractive summarization or topic modeling combined, whereas in the general domain, topic modeling and extractive summarization techniques are the most researched [58,159]. This discrepancy suggests that despite abstractive summarization techniques being immature [160], general problems with extractive summarization, such as redundancy [161], lack of coherence [162,163], and lengthiness [163], can be problematic for clinical applications. The verdict about topic modeling is unclear. Arnold et al [112] argue that clinicians are good at interpreting topic model results, but other records using this technique do not present evaluations with human participants.

An alternative (and natural) way of organizing summarization methods is to assess how they contribute to the summarization process. Motivated by the lack of a widely accepted theory of the summarization process, this review proposes a 3-step (collect—synthesize—communicate) framework to describe the summarization process, where each step should ideally be addressed by all summarization methods.

For the information collection step, many studies assume an easily queriable information source or propose medical concept extraction from textual data as a solution. More complex information (context, syntactic, or semantic structure of statements) is extracted in only a few studies, and some works propose extracting specific aspects as information.

Concerning information synthesis, a common approach is to precisely define the content of the summary (eg, [118,165,166]) or at least its format [167]. However, these studies do not evaluate the quality of their proposition (except the study by Ham et al [119]). In contrast, some records carried out experiments on the information needs of physicians [158,168]; however, the results were not integrated into any of the reviewed systems.

Concerning summary visualizations, there is no clear opinion on whether textual or graphical summaries are preferable in the medical context. Although there is a slight dominance of graphical displays among the analyzed records, some works [169,170] argue that textual summaries lead to more accurate decisions. However, a general pattern of these works is that they compare a specific graphical display with a particular textual display, limiting generalizability. These contradictory results suggest that both formats are helpful for clinical summarization, if relevant features are present. Problem-oriented views presented in some records (eg, [120]) can include both types of display and might have other advantages, as they group all available information about patient-specific problems [171].

Concerning the view on summarization, both top-down and bottom-up approaches are justifiable in a clinical setting. However, several bottom-up approaches have been inspired by studies that use top-down approaches. One example is the recent development of techniques for identifying salient concepts in source documents for abstractive summarization. This phenomenon may be due to the natural need for accountability and interpretability, which can be achieved more easily with a bottom-up approach closer to human cognition. The need for bottom-up approaches also suggests that there is a need that summarization techniques address all 3 steps of the proposed "collect—synthesize—communicate" framework, including information collection, synthesis, and visualization.

How Are the Temporal, Uncertain Aspects of Information and Its Medical Persistence Conserved and Used?

The temporal nature of clinical data is an essential aspect of clinical reasoning [172], and a relatively large portion of analyzed records presents solutions to use this aspect of information. However, in most of these studies, this aspect was only represented as a visualization feature. Most visualizations are timeline visualizations, plotting information along a horizontal or vertical temporal axis following Plaisant et al [175], although some alternative methods exist [121-124]. Alternatively, some studies have revealed the relationship between events by analyzing how variables change during the patient journey.

Although temporality in clinical settings is believed to be more complex [172] than a series of punctual events, current solutions to clinical summarization hardly reflect this complexity. Very few studies have attempted to incorporate more temporal information by using more complex models of temporality (eg, events lasting during an interval). Complex temporal information is usually not directly available in patient records and must be deduced from the context and knowledge-based rules. This process is called temporal abstraction and was applied in previous studies [90,91,129-132,176]. Hunter et al [129,130] considered the uncertainty of temporal information by defining the beginning and end of each time interval as an interval.

Although several levels of uncertainty exist in clinical care [177], the majority of the analyzed reports do not present solutions to conserve or handle any information uncertainty.

To a lesser extent, the pertinence of medical knowledge has been overlooked by many summarization approaches. Many records do not consider medical pertinence or use it only for some design considerations. However, the few records that handle this aspect of medical data provide a relatively wide range of solutions to constrain the resulting summaries. In most cases, these constraints are relatively weak; for example, concepts are assumed to be part of a specific medical ontology. This is obviously the case for concept extraction tools, but the records using reinforcement learning approximate factual correctness using this approach as well.

Deeper integration of medical knowledge is only present in works using medical rules to select salient information and in the 2 works using medical rules to create summaries. Liang et al [9] used medical knowledge to create components of their proposed NLP pipeline, whereas Shi et al [102] used medical knowledge–based rules to visualize abnormalities in the human body.

How to Identify a Good Summary?

Using the definition of clinical summarization (ie, simplifying and presenting information so that HCPs can act more smoothly and efficiently in different clinical situations), the ultimate purpose of an evaluation might be to determine whether using the proposed summarization systems would improve the efficiency of HCPs.

However, such an evaluation is often unfeasible owing to the high costs and ethical issues associated with potential medical errors.

This is supported by the results, as many of the proposed evaluations are approximative solutions, and there is quasi-uniform agreement that evaluating summarization is challenging and suboptimal [168]. The spectrum of these evaluations is broad, but a common trend is to carry out a qualitative evaluation using an easily calculable evaluation metric describing either the quality of the summary or its "usefulness" to perform a proxy task (ie, disease prediction).

These qualitative analyses are suboptimal. For example, one of the most common qualitative metrics, the ROUGE score, assumes a human-annotated "gold standard" summary to which to compare, but this standard may not exist given the high cost of annotation or because there are disagreements between people about what would be a "gold standard summary" [70,168]. To tackle this problem, some records [72,178] present a comparison between the semantic distribution of the input and the summary, whereas others [133,150,179] use heuristics to evaluate the results. Another problem with the ROUGE score is that even with a high ROUGE score, a summary can be very inaccurate [168]; therefore, there have been attempts to measure the "faithfulness" of summaries either by the number of medical concepts retrieved [81] or with a more complex faithfulness measure defined by Zhang et al [83].

Evaluations with human participants often complement the qualitative evaluations. Human evaluations have mainly positive outcomes (except in the study by van Amsterdam et al [148]); however, most of the evaluations are carried out on a small scale. This can explain why very few long-lasting implementations in health care have been presented in the literature.

It is also important to mention that there is very little comparison between summarization methods, and citations between records are scarce. This suggests that the research in this domain is fragmented.

These shortcomings suggest that evaluation is a weak point in clinical summarization proposals, and the lack of widely accepted evaluation metrics and methodology might be a main obstacle for research in the field.

Limitations

Methodological biases are present in selection, synthesis, and reporting. First, the number of reviewers was limited both in the selection and analysis of records, resulting in selection and synthesis bias.

Selection bias also comes from the fact that the review was carried out on works published in a scientific paper or in the gray literature, and the initial search was carried out on 2 databases that are more specific to medicine. However, several unpublished summarization solutions have been applied to current EHR systems.

Moreover, publishing bias also adds to selection bias, as there is a clear dominance of positive results in scientific publishing.

Furthermore, the applied research queries and the selection of the database are also a source of bias. Using other data sources (eg, IEEE Xplore and Scopus) might have introduced further bias to the analysis. However, including the citations and references in the review process might have reduced this bias significantly. Moreover, the queries are formulated in English; therefore, results not in English and containing non-English terms might be missed if their abstract was not translated or if they do not appear in the citation list or references in the retrieved articles. Finally, there were potentially relevant records [149,180,181], where the full text could not be read and analyzed as it was not available at the time of writing the manuscript.

Conclusions

Clinical summarization has not reached all domains at the same pace, although it is potentially beneficial in most medical fields. Two aspects, the requirement for quick decision-making and the overabundance of data, were identified as the main drivers for the development of automatic summarization methods. However, other less-evident drivers might also play a role in adaptation.

Despite this need, very few [113,119,182] scientific publications are presenting adaptation in real clinical settings, suggesting a low success rate in clinical environments.

Despite the large number and variety of propositions, hardly any comparisons exist between the solutions. This low rate is due to the difficulty in comparing the summarization methods. From a cognitive psychological perspective and to measure how the summarization methods align with the definition of summarization, this review proposes to compare these algorithms through a "collect—synthesis—communicate" framework referring to information gathering from data, its synthesis, and communication to the end user.

Only a small proportion of the current propositions address all 3 steps, and none of the most abundant methods (ie, abstractive, extractive summarization, and visual design) address them completely.

Beyond the lack of alignment of the dimensions of summarization, propositions conserve and use crucial aspects of information (temporality, uncertainty, and medical pertinence) to varying extents.

Although uncertainty is rarely considered, temporality and some medical pertinence are conserved during some presentations, but the solutions are often preliminary or lack depth in these aspects. Further research is necessary to address these issues.

Nevertheless, the main shortcoming of the current automatic summarization methods is the lack of consistent evaluation. Although there are some new proposals to evaluate the quality of summarizations more rigorously [83], further research is required to relate these metrics to human perceptions.

Acknowledgments

This work has been supported by "The Fondation privée des HUG" and the University Hospitals of Geneva. The authors would like to thank the reviewers for their valuable comments, which helped them improve the quality of the manuscript.

Conflicts of Interest

CL is the editor-in-chief of the JMI Journal, but was not involved in the process of reviewing/accepting this paper. The author have no further interests to declare.

Multimedia Appendix 1
PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist for the review.
[PDF File (Adobe PDF File), 497 KB - medinform_v11i1e44639_app1.pdf]

Multimedia Appendix 2
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 flow diagram describing the review process applied in the study [182].
[DOCX File , 59 KB - medinform_v11i1e44639_app2.docx]

Multimedia Appendix 3
Histogram showing the distribution of the number of evaluators in studies where evaluations with human participants are present.
[DOCX File , 53 KB - medinform_v11i1e44639_app3.docx]

Multimedia Appendix 4
Histogram showing the number of other publications appearing in the review cited by each report. The number of cross-citations between the works are low.
[DOCX File , 50 KB - medinform_v11i1e44639_app4.docx]

References

1. Alterman R. Understanding and summarization. Artif Intell Rev 1991;5(4):239-254 [FREE Full text] [doi: 10.1007/bf00141756]
2. Feblowitz JC, Wright A, Singh H, Samal L, Sittig DF. Summarization of clinical information: a conceptual model. J Biomed Inform 2011 Aug;44(4):688-699 [FREE Full text] [doi: 10.1016/j.jbi.2011.03.008] [Medline: 21440086]
3. Powsner SM, Tufte ER. Graphical summary of patient status. Lancet 1994 Aug 06;344(8919):386-389. [doi: 10.1016/s0140-6736(94)91406-0] [Medline: 7914312]
4. Hall A, Walton G. Information overload within the health care system: a literature review. Health Info Libr J 2004 Jun;21(2):102-108 [FREE Full text] [doi: 10.1111/j.1471-1842.2004.00506.x] [Medline: 15191601]
5. Botsis T, Hartvigsen G, Chen F, Weng C. Secondary use of EHR: data quality issues and informatics opportunities. Summit Transl Bioinform 2010 Mar 01;2010:1-5 [FREE Full text] [Medline: 21347133]
6. Nijor S, Rallis G, Lad N, Gokcen E. Patient safety issues from information overload in electronic medical records. J Patient Saf 2022 Sep 01;18(6):e999-1003 [FREE Full text] [doi: 10.1097/PTS.0000000000001002] [Medline: 35985047]
7. Mamykina L, Vawdrey DK, Stetson PD, Zheng K, Hripcsak G. Clinical documentation: composition or synthesis? J Am Med Inform Assoc 2012 Nov;19(6):1025-1031 [FREE Full text] [doi: 10.1136/amiajnl-2012-000901] [Medline: 22813762]
8. O'Donnell HC, Kaushal R, Barrón Y, Callahan MA, Adelman RD, Siegler EL. Physicians' attitudes towards copy and pasting in electronic note writing. J Gen Intern Med 2009 Jan 8;24(1):63-68 [FREE Full text] [doi: 10.1007/s11606-008-0843-2] [Medline: 18998191]
9. Liang JJ, Tsou CH, Dandala B, Poddar A, Joopudi V, Mahajan D, et al. Reducing physicians' cognitive load during chart review: a problem-oriented summary of the patient electronic record. AMIA Annu Symp Proc 2022 Feb 21;2021:763-772 [FREE Full text] [Medline: 35308927]
10. Pivovarov R, Elhadad N. Automated methods for the summarization of electronic health records. J Am Med Inform Assoc 2015 Sep;22(5):938-947 [FREE Full text] [doi: 10.1093/jamia/ocv032] [Medline: 25882031]
11. Mishra R, Bian J, Fiszman M, Weir C, Jonnalagadda S, Mostafa J, et al. Text summarization in the biomedical domain: a systematic review of recent research. J Biomed Inform 2014 Dec;52:457-467 [FREE Full text] [doi: 10.1016/j.jbi.2014.06.009] [Medline: 25016293]
12. Wang M, Wang M, Yu F, Yang Y, Walker J, Mostafa J. A systematic review of automatic text summarization for biomedical literature and EHRs. J Am Med Inform Assoc 2021 Sep 18;28(10):2287-2297 [FREE Full text] [doi: 10.1093/jamia/ocab143] [Medline: 34338801]
13. Afantenos S, Karkaletsis V, Stamatopoulos P. Summarization from medical documents: a survey. Artif Intell Med 2005 Feb;33(2):157-177. [doi: 10.1016/j.artmed.2004.07.017] [Medline: 15811783]
14. Li I, Pan J, Goldwasser J, Verma N, Wong WP, Nuzumlalı MY, et al. Neural natural language processing for unstructured data in electronic health records: a review. Comput Sci Rev 2022 Nov;46:100511 [FREE Full text] [doi: 10.1016/j.cosrev.2022.100511]
15. Koleck TA, Dreisbach C, Bourne PE, Bakken S. Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. J Am Med Inform Assoc 2019 Apr 01;26(4):364-379 [FREE Full text] [doi: 10.1093/jamia/ocy173] [Medline: 30726935]
16. van Buchem MM, Boosman H, Bauer MP, Kant IM, Cammel SA, Steyerberg EW. The digital scribe in clinical practice: a scoping review and research agenda. NPJ Digit Med 2021 Mar 26;4(1):57 [FREE Full text] [doi: 10.1038/s41746-021-00432-5] [Medline: 33772070]
17. Boyd AD, Young CD, Amatayakul M, Dieter MG, Pawola LM. Developing visual thinking in the electronic health record. Stud Health Technol Inform 2017;245:308-312. [Medline: 29295105]
18. Rind A, Federico P, Gschwandtner T, Aigner W, Doppler J, Wagner M. Visual analytics of electronic health records with a focus on time. In: Rinaldi G, editor. New Perspectives in Medical Records: Meeting the Needs of Patients and Practitioners. Cham, Switzerland: Springer; 2017:65-77.
19. West VL, Borland D, Hammond WE. Innovative information visualization of electronic health record data: a systematic review. J Am Med Inform Assoc 2015 Mar;22(2):330-339 [FREE Full text] [doi: 10.1136/amiajnl-2014-002955] [Medline: 25336597]
20. Dowding D, Randell R, Gardner P, Fitzpatrick G, Dykes P, Favela J, et al. Dashboards for improving patient care: review of the literature. Int J Med Inform 2015 Feb;84(2):87-100 [FREE Full text] [doi: 10.1016/j.ijmedinf.2014.10.001] [Medline: 25453274]
21. Wright MC, Borbolla D, Waller RG, Del Fiol G, Reese T, Nesbitt P, et al. Critical care information display approaches and design frameworks: a systematic review and meta-analysis. J Biomed Inform X 2019 Sep;3:100041 [FREE Full text] [doi: 10.1016/j.yjbinx.2019.100041] [Medline: 31423485]
22. Segall N, Borbolla D, Del FG, Waller R, Reese T, Nesbitt P, et al. Trend displays to support critical care: a systematic review. In: Proceedings of the 2017 IEEE International Conference on Healthcare Informatics. 2017 Presented at: ICHI '17; August 23-26, 2017; Park City, UT p. 305-313 URL: https://ieeexplore.ieee.org/abstract/document/8031160

23. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
24. Kellermeyer L, Harnke B, Knight S. Covidence and rayyan. J Med Libr Assoc 2018 Oct 04;106(4):580-583 [FREE Full text] [doi: 10.5195/jmla.2018.513]
25. Nelson KE. What do you do if you can't tell the whole story? The development of summarization skills. In: Nelson KE, editor. Children's Language. Volume 4. New York, NY: Psychology Press; 1983:315-383.
26. Hidi S, Anderson V. Producing written summaries: task demands, cognitive operations, and implications for instruction. Rev Educ Res 1986;56(4):473-493 [FREE Full text] [doi: 10.3102/00346543056004473]
27. Spärck Jones K. Automatic summarising: the state of the art. Inf Process Manag 2007 Nov;43(6):1449-1481 [FREE Full text] [doi: 10.1016/j.ipm.2007.03.009]
28. Jones K. Automatic summarizing: factors and directions. In: Mani I, Maybury MT, editors. Advances in Automatic Text Summarization. Cambridge, MA: MIT Press; 1999:1-12.
29. Liang J, Tsou C, Poddar A. A novel system for extractive clinical note summarization using EHR data. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop. 2019 Presented at: ClinicalNLP '19; June 5-7, 2019; Minneapolis, MN p. 46-54 URL: https://aclanthology.org/W19-1906.pdf [doi: 10.18653/v1/w19-1906]
30. Gundogdu B, Pamuksuz U, Chung JH, Telleria JM, Liu P, Khan F, et al. Customized impression prediction from radiology reports using BERT and LSTMs. IEEE Trans Artif Intell 2023 Aug;4(4):744-753. [doi: 10.1109/tai.2021.3086435]
31. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
32. Abacha AB, Mrabet Y, Zhang Y, Shivade C, Langlotz C, Demner-Fushman D. Overview of the MEDIQA 2021 shared task on summarization in the medical domain. In: Proceedings of the 20th Workshop on Biomedical Language Processing. 2021 Presented at: BioNLP '21; June 11, 2021; Virtual Event p. 74-85 URL: https://aclanthology.org/2021.bionlp-1.8.pdf [doi: 10.18653/v1/2021.bionlp-1.8]
33. Moen H, Peltonen LM, Heimonen J, Airola A, Pahikkala T, Salakoski T, et al. Comparison of automatic summarisation methods for clinical free text notes. Artif Intell Med 2016 Feb;67:25-37 [FREE Full text] [doi: 10.1016/j.artmed.2016.01.003] [Medline: 26900011]
34. Reunamo A, Peltonen P, Mustonen M, Saari M, Salakoski T, Salanterä S, et al. Text classification model explainability for keyword extraction - towards keyword-based summarization of nursing care episodes. Stud Health Technol Inform 2022 Jun 06;290:632-636 [FREE Full text] [doi: 10.3233/SHTI220154] [Medline: 35673093]
35. Deng Y, Denecke K. Visualizing unstructured patient data for assessing diagnostic and therapeutic history. Stud Health Technol Inform 2014;205:1158-1162. [Medline: 25160371]
36. Liang S, Kades K, Fink M, Full P, Weber T, Kleesiek J, et al. Fine-tuning BERT models for summarizing German radiology findings. In: Proceedings of the 4th Clinical Natural Language Processing Workshop. 2022 Presented at: ClinicalNLP '22; July 14, 2022; Virtual Event p. 30-40 URL: https://aclanthology.org/2022.clinicalnlp-1.4.pdf [doi: 10.18653/v1/2022.clinicalnlp-1.4]
37. Loukas C. Video content analysis of surgical procedures. Surg Endosc 2018 Feb 26;32(2):553-568. [doi: 10.1007/s00464-017-5878-1] [Medline: 29075965]
38. Koopman RJ, Mainous AG3. Evaluating multivariate risk scores for clinical decision making. Fam Med 2008 Jun;40(6):412-416. [Medline: 18773779]
39. Miller RL, Vergara VM, Pearlson GD, Calhoun VD. Multiframe Evolving Dynamic Functional Connectivity (EVOdFNC): a method for constructing and investigating functional brain motifs. Front Neurosci 2022 Apr 19;16:770468 [FREE Full text] [doi: 10.3389/fnins.2022.770468] [Medline: 35516809]
40. Hamza R, Muhammad K, Lv Z, Titouna F. Secure video summarization framework for personalized wireless capsule endoscopy. Pervasive Mob Comput 2017 Oct;41:436-450 [FREE Full text] [doi: 10.1016/j.pmcj.2017.03.011]
41. Muhammad K, Ahmad J, Sajjad M, Baik SW. Visual saliency models for summarization of diagnostic hysteroscopy videos in healthcare systems. Springerplus 2016 Sep 6;5(1):1495 [FREE Full text] [doi: 10.1186/s40064-016-3171-8] [Medline: 27652068]
42. Ma M, Mei S, Wan S, Wang Z, Ge Z, Lam V, et al. Keyframe extraction from laparoscopic videos via diverse and weighted dictionary selection. IEEE J Biomed Health Inform 2021 May;25(5):1686-1698. [doi: 10.1109/JBHI.2020.3019198] [Medline: 32841131]
43. Yaesoubi M, Silva RF, Iraji A, Calhoun VD. Frequency-aware summarization of resting-state fMRI data. Front Syst Neurosci 2020 Apr 7;14:16 [FREE Full text] [doi: 10.3389/fnsys.2020.00016] [Medline: 32317942]
44. Huang R, Ying Q, Lin Z, Zheng Z, Tan L, Tang G, et al. Extracting keyframes of breast ultrasound video using deep reinforcement learning. Med Image Anal 2022 Aug;80:102490. [doi: 10.1016/j.media.2022.102490] [Medline: 35717873]
45. Hendryx EP, Rivière BM, Sorensen DC, Rusin CG. Finding representative electrocardiogram beat morphologies with CUR. J Biomed Inform 2018 Jan;77:97-110 [FREE Full text] [doi: 10.1016/j.jbi.2017.12.003] [Medline: 29224855]

46. Park J, Kang K. HeartSearcher: finds patients with similar arrhythmias based on heartbeat classification. IET Syst Biol 2015 Dec;9(6):303-308 [FREE Full text] [doi: 10.1049/iet-syb.2015.0011] [Medline: 26577165]
47. Akalin A, Franke V, Vlahoviček K, Mason CE, Schübeler D. Genomation: a toolkit to summarize, annotate and visualize genomic intervals. Bioinformatics 2015 Apr 01;31(7):1127-1129. [doi: 10.1093/bioinformatics/btu775] [Medline: 25417204]
48. Cai X, Liu S, Han J, Yang L, Liu Z, Liu T. ChestXRayBERT: a pretrained language model for chest radiology report summarization. IEEE Trans Multimedia 2021;25:845-855 [FREE Full text] [doi: 10.1109/tmm.2021.3132724]
49. MacAvaney S, Sotudeh S, Cohan A, Goharian N, Talati I, Filice R. Ontology-aware clinical abstractive summarization. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2019 Presented at: SIGIR'19; July 21-25, 2019; Paris, France p. 1013-1016 URL: https://dl.acm.org/doi/10.1145/3331184.3331319 [doi: 10.1145/3331184.3331319]
50. Sotudeh S, Goharian N, Filice R. Attend to medical ontologies: content selection for clinical abstractive summarization. arXiv. Preprint posted online May 1, 2020 2020 [FREE Full text] [doi: 10.18653/v1/2020.acl-main.172]
51. Zhang Y, Ding D, Qian T, Manning C, Langlotz C. Learning to summarize radiology findings. arXiv. Preprint posted online September 12, 2018 2018 [FREE Full text] [doi: 10.18653/v1/w18-5623]
52. Manas G, Aribandi V, Kursuncu U, Alambo A, Shalin VL, Thirunarayan K, et al. Knowledge-infused abstractive summarization of clinical diagnostic interviews: framework development study. JMIR Ment Health 2021 May 10;8(5):e20865 [FREE Full text] [doi: 10.2196/20865] [Medline: 33970116]
53. Chen YP, Chen YY, Lin JJ, Huang CH, Lai F. Modified bidirectional encoder representations from transformers extractive summarization model for hospital information systems based on character-level tokens (AlphaBERT): development and performance evaluation. JMIR Med Inform 2020 Apr 29;8(4):e17787 [FREE Full text] [doi: 10.2196/17787] [Medline: 32347806]
54. Gaut G, Steyvers M, Imel ZE, Atkins DC, Smyth P. Content coding of psychotherapy transcripts using labeled topic models. IEEE J Biomed Health Inform 2017 Mar;21(2):476-487 [FREE Full text] [doi: 10.1109/JBHI.2015.2503985] [Medline: 26625437]
55. Speier W, Ong MK, Arnold CW. Using phrases and document metadata to improve topic modeling of clinical reports. J Biomed Inform 2016 Jun;61:260-266 [FREE Full text] [doi: 10.1016/j.jbi.2016.04.005] [Medline: 27109931]
56. Shah J, Mohammed S. Clinical narrative summarization based on the MIMIC III dataset. J Multimedia Ubiquitous Eng 2020 Nov;15(2):60 [FREE Full text] [doi: 10.21742/IJMUE.2020.15.2.05]
57. Chen JH, Goldstein MK, Asch SM, Mackey L, Altman RB. Predicting inpatient clinical order patterns with probabilistic topic models vs conventional order sets. J Am Med Inform Assoc 2017 May 01;24(3):472-480 [FREE Full text] [doi: 10.1093/jamia/ocw136] [Medline: 27655861]
58. Allahyari M, Pouriyeh S, Assefi M, Safaei S, D. ED, B. JB, et al. Text summarization techniques: a brief survey. Int J Adv Comput Sci Appl 2017;8(10):397-405 [FREE Full text] [doi: 10.14569/ijacsa.2017.081052]
59. Jane Y, Nehemiah HK, Kannan A. Classifying unevenly spaced clinical time series data using forecast error approximation based bottom-up (FeAB) segmented time delay neural network. Comput Methods Biomech Biomed Eng Imaging Vis 2020 Dec 21;9(1):92-105. [doi: 10.1080/21681163.2020.1817791]
60. Portet F, Reiter E, Gatt A, Hunter J, Sripada S, Freer Y, et al. Automatic generation of textual summaries from neonatal intensive care data. Artif Intell 2009 May;173(7-8):789-816 [FREE Full text] [doi: 10.1016/j.artint.2008.12.002]
61. Hsu W, Taira RK. Tools for improving the characterization and visualization of changes in neuro-oncology patients. AMIA Annu Symp Proc 2010 Nov 13;2010:316-320 [FREE Full text] [Medline: 21346992]
62. Albers D, Elhadad N, Claassen J, Perotte R, Goldstein A, Hripcsak G. Estimating summary statistics for electronic health record laboratory data for use in high-throughput phenotyping algorithms. J Biomed Inform 2018 Feb;78:87-101 [FREE Full text] [doi: 10.1016/j.jbi.2018.01.004] [Medline: 29369797]
63. Landi I, Glicksberg BS, Lee H, Cherng S, Landi G, Danieletto M, et al. Deep representation learning of electronic health records to unlock patient stratification at scale. NPJ Digit Med 2020 Jul 17;3(1):96 [FREE Full text] [doi: 10.1038/s41746-020-0301-z] [Medline: 32699826]
64. Miotto R, Li L, Kidd B, Dudley J. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. Sci Rep 2016 May 17;6:26094 [FREE Full text] [doi: 10.1038/srep26094] [Medline: 27185194]
65. Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, et al. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv. Preprint posted online October 29, 2019 2019 [FREE Full text] [doi: 10.18653/v1/2020.acl-main.703]
66. Devlin J, Chang M, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv. Preprint posted online October 11, 2018 2018 [FREE Full text]
67. Zhang J, Zhao Y, Saleh M, Liu P. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. In: Proceedings of the 37th International Conference on Machine Learning. 2020 Presented at: ICML '20; July 13-18, 2020; Vienna, Austria p. 11328-11339 URL: https://dl.acm.org/doi/abs/10.5555/3524938.3525989 [doi: 10.5555/3524938.3525989]
68. Ribeiro M, Singh S, Guestrin C. "Why should I trust you?": Explaining the predictions of any classifier. arXiv. Preprint posted online February 16, 2016 2016 Aug 9 [FREE Full text] [doi: 10.1145/2939672.2939778]

69. Subramanian V, Engelhard M, Berchuck S, Chen L, Henao R, Carin L. SpanPredict: extraction of predictive document spans with neural attention. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language. 2021 Presented at: NAACL '21; June 6-11, 2021; Stroudsburg, PA p. 5234-5258 URL: https://aclanthology.org/2021.naacl-main.413.pdf [doi: 10.18653/v1/2021.naacl-main.413]
70. Alsentzer E, Kim A. Extractive summarization of ehr discharge notes. arXiv. Preprint posted online October 26, 2018 2018 [FREE Full text]
71. Liu X, Xu K, Xie P, Xing E. Unsupervised pseudo-labeling for extractive summarization on electronic health records. arXiv. Preprint posted online November 20, 2018 2018 [FREE Full text]
72. Kanwal N, Rizzo G. Attention-based clinical note summarization. In: Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing. 2022 Presented at: SAC '22; April 25-29, 2022; Virtual Event p. 813-820 URL: https://dl.acm.org/doi/abs/10.1145/3477314.3507256 [doi: 10.1145/3477314.3507256]
73. McInerney DJ, Dabiri B, Touret AS, Young G, Meent JW, Wallace B. Query-focused EHR summarization to aid imaging diagnosis. Proc Mach Learn Res 2020:632-659 [FREE Full text]
74. Hu J, Li J, Chen Z, Shen Y, Song Y, Wan X. Word graph guided summarization for radiology findings. arXiv. Preprint posted online December 18, 2021 2021 [FREE Full text] [doi: 10.18653/v1/2021.findings-acl.441]
75. Hu J, Li Z, Chen Z, Li Z, Wan X. Graph enhanced contrastive learning for radiology findings summarization. arXiv. Preprint posted online April 1, 2022 2022 [FREE Full text] [doi: 10.18653/v1/2022.acl-long.320]
76. Chang D, Lin E, Brandt C, Taylor RA. Incorporating domain knowledge into language models by using graph convolutional networks for assessing semantic textual similarity: model development and performance comparison. JMIR Med Inform 2021 Nov 26;9(11):e23101 [FREE Full text] [doi: 10.2196/23101] [Medline: 34842531]
77. Mahajan D, Tsou CH, Liang JL. IBMResearch at MEDIQA 2021: toward improving factual correctness of radiology report abstractive summarization. In: Proceedings of the 20th Workshop on Biomedical Language Processing. 2021 Presented at: BioNLP '21; June 11, 2021; Virtual Event p. 302-310 URL: https://aclanthology.org/2021.bionlp-1.35.pdf [doi: 10.18653/v1/2021.bionlp-1.35]
78. Dai S, Wang Q, Lyu Y, Zhu Y. BDKG at MEDIQA 2021: system report for the radiology report summarization task. In: Proceedings of the 20th Workshop on Biomedical Language Processing. 2021 Presented at: BioNLP '21; June 11, 2021; Virtual Event p. 103-111 URL: https://aclanthology.org/2021.bionlp-1.11.pdf [doi: 10.18653/v1/2021.bionlp-1.11]
79. He Y, Chen M, Huang S. damo_nlp at MEDIQA 2021: knowledge-based preprocessing and coverage-oriented reranking for medical question summarization. In: Proceedings of the 20th Workshop on Biomedical Language Processing. 2021 Presented at: BioNLP '21; June 11, 2021; Virtual Event p. 2112-2118 URL: https://aclanthology.org/2021.bionlp-1.12.pdf [doi: 10.18653/v1/2021.bionlp-1.12]
80. Kondadadi R, Manch S, Ngo J, McCormack R. Optum at MEDIQA 2021: abstractive summarization of radiology reports using simple BART finetuning. In: Proceedings of the 20th Workshop on Biomedical Language Processing. 2021 Presented at: BioNLP '20; June 11, 2021; Virtual Event p. 280-284 URL: https://aclanthology.org/2021.bionlp-1.32.pdf [doi: 10.18653/v1/2021.bionlp-1.32]
81. Shing H, Shivade C, Pourdamghani N, Nan F, Resnik P, Oard D, et al. Towards clinical encounter summarization: learning to compose discharge summaries from prior notes. arXiv. Preprint posted online April 27, 2021 2021 [FREE Full text]
82. Xu L, Zhang Y, Hong L, Cai Y, Sung S. ChicHealth@ MEDIQA 2021xploring the limits of pre-trained seq2seq models for medical summarization. In: Proceedings of the 20th Workshop on Biomedical Language Processing. 2021 Presented at: BioNLP '21; June 11, 2021; Virtual Event p. 263-267 URL: https://aclanthology.org/2021.bionlp-1.29.pdf [doi: 10.18653/v1/2021.bionlp-1.29]
83. Zhang Y, Merck D, Tsai EB, Manning CD, Langlotz CP. Optimizing the factual correctness of a summary: a study of summarizing radiology reports. arXiv. Preprint posted online November 6, 2019 2019 [FREE Full text] [doi: 10.18653/v1/2020.acl-main.458]
84. Delbrouck J, Zhang C, Rubin D. QIAI at MEDIQA 2021: multimodal radiology report summarization. In: Proceedings of the 20th Workshop on Biomedical Language Processing. 2021 Presented at: BioNLP '21; June 11, 2021; Virtual Event p. 285-290 URL: https://aclanthology.org/2021.bionlp-1.33.pdf [doi: 10.18653/v1/2021.bionlp-1.33]
85. Xu B, Gil-Jardiné C, Thiessard F, Tellier E, Avalos-Fernandez M, Lagarde E. Pre-training a neural language model improves the sample efficiency of an emergency room classification model. arXiv. Preprint posted online August 30, 2019 2019 [FREE Full text]
86. Tang L, Kooragayalu S, Wang Y, Ding Y, Durrett G, Rousseau JF, et al. EchoGen: a new benchmark study on generating conclusions from echocardiogram notes. Proc Conf Assoc Comput Linguist Meet 2022 May;2022:359-368 [FREE Full text] [doi: 10.18653/v1/2022.bionlp-1.35] [Medline: 36339656]
87. Lee E, Uppal K. CERC: an interactive content extraction, recognition, and construction tool for clinical and biomedical text. BMC Med Inform Decis Mak 2020 Dec 15;20(Suppl 14):306 [FREE Full text] [doi: 10.1186/s12911-020-01330-8] [Medline: 33323109]
88. Devarakonda MV, Mehta N, Tsou C, Liang JJ, Nowacki AS, Jelovsek JE. Automated problem list generation and physicians perspective from a pilot study. Int J Med Inform 2017 Sep;105:121-129 [FREE Full text] [doi: 10.1016/j.ijmedinf.2017.05.015] [Medline: 28750905]

89. Zhu W, He Y, Chai L, Fan Y, Ni Y, Xie G, et al. paht_nlp @ MEDIQA 2021: multi-grained query focused multi-answer summarization. In: Proceedings of the 20th Workshop on Biomedical Language Processing. 2021 Presented at: BioNLP '21; June 11, 2021; Virtual Event p. 966 URL: https://aclanthology.org/2021.bionlp-1.10.pdf [doi: 10.18653/v1/2021.bionlp-1.10]
90. Goldstein A, Shahar Y. An automated knowledge-based textual summarization system for longitudinal, multivariate clinical data. J Biomed Inform 2016 Jun;61:159-175 [FREE Full text] [doi: 10.1016/j.jbi.2016.03.022] [Medline: 27039119]
91. Goldstein A, Shahar Y, Orenbuch E, Cohen MJ. Evaluation of an automated knowledge-based textual summarization system for longitudinal clinical data, in the intensive care domain. Artif Intell Med 2017 Oct;82:20-33. [doi: 10.1016/j.artmed.2017.09.001] [Medline: 28958803]
92. Van Vleck TT, Elhadad N. Corpus-based problem selection for EHR note summarization. AMIA Annu Symp Proc 2010 Nov 13;2010:817-821 [FREE Full text] [Medline: 21347092]
93. Müller H, Reihs R, Posch A, Kremer A, Ulrich D, Zatloukal K. Data driven GUI design and visualization for a NGS based clinical decision support system. In: Proceedings of the 2016 20th International Conference Information Visualisation. 2016 Presented at: IV '16; July 19-22, 2016; Lisbon, Portugal p. 355-360 URL: https://ieeexplore.ieee.org/abstract/document/7557952 [doi: 10.1109/iv.2016.79]
94. Goff DJ, Loehfelm TW. Automated radiology report summarization using an open-source natural language processing pipeline. J Digit Imaging 2018 Apr;31(2):185-192 [FREE Full text] [doi: 10.1007/s10278-017-0030-2] [Medline: 29086081]
95. Kim B, Merchant M, Zheng C, Thomas A, Contreras R, Jacobsen S, et al. A natural language processing program effectively extracts key pathologic findings from radical prostatectomy reports. J Endourol 2014 Dec;28(12):1474-1478. [doi: 10.1089/end.2014.0221] [Medline: 25211697]
96. Moradi M, Ghadiri N. Different approaches for identifying important concepts in probabilistic biomedical text summarization. Artif Intell Med 2018 Jan;84:101-116. [doi: 10.1016/j.artmed.2017.11.004] [Medline: 29208328]
97. Mane K, Bizon C, Schmitt C, Owen P, Burchett B, Pietrobon R, et al. VisualDecisionLinc: a visual analytics approach for comparative effectiveness-based clinical decision support in psychiatry. J Biomed Inform 2012 Feb;45(1):101-106 [FREE Full text] [doi: 10.1016/j.jbi.2011.09.003] [Medline: 21963813]
98. Mathe JL, Martin JB, Miller P, Ledeczi A, Weavind LM, Nadas A, et al. A model-integrated, guideline-driven, clinical decision-support system. IEEE Softw 2009 Jul;26(4):54-61 [FREE Full text] [doi: 10.1109/ms.2009.84]
99. Semler M, Weavind L, Hooper M, Rice T, Gowda S, Nadas A, et al. An electronic tool for the evaluation and treatment of sepsis in the ICU: a randomized controlled trial. Crit Care Med 2015 Aug;43(8):1595-1602 [FREE Full text] [doi: 10.1097/CCM.0000000000001020] [Medline: 25867906]
100. Tignanelli CJ, Silverman GM, Lindemann EA, Trembley AL, Gipson JC, Beilman G, et al. Natural language processing of prehospital emergency medical services trauma records allows for automated characterization of treatment appropriateness. J Trauma Acute Care Surg 2020 May;88(5):607-614 [FREE Full text] [doi: 10.1097/TA.0000000000002598] [Medline: 31977990]
101. Klumpner TT, Kountanis JA, Langen ES, Smith RD, Tremper KK. Use of a novel electronic maternal surveillance system to generate automated alerts on the labor and delivery unit. BMC Anesthesiol 2018 Jun 26;18(1):78 [FREE Full text] [doi: 10.1186/s12871-018-0540-6] [Medline: 29945569]
102. Shi L, Sun J, Yang Y, Ling T, Wang M, Gu Y, et al. Three-dimensional visual patient based on electronic medical diagnostic records. IEEE J Biomed Health Inform 2018 Jan;22(1):161-172. [doi: 10.1109/JBHI.2017.2702201] [Medline: 28500014]
103. Hsu W, Taira R, El-Saden S, Kangarloo H, Bui A. Context-based electronic health record: toward patient specific healthcare. IEEE Trans Inf Technol Biomed 2012 Mar;16(2):228-234 [FREE Full text] [doi: 10.1109/TITB.2012.2186149] [Medline: 22395637]
104. Ahmed A, Chandra S, Herasevich V, Gajic O, Pickering BW. The effect of two different electronic health record user interfaces on intensive care provider task load, errors of cognition, and performance. Crit Care Med 2011 Jul;39(7):1626-1634. [doi: 10.1097/CCM.0b013e31821858a0] [Medline: 21478739]
105. Foraker R, Kite B, Kelley M, Lai A, Roth C, Lopetegui M, et al. EHR-based visualization tool: adoption rates, satisfaction, and patient outcomes. EGEMS (Wash DC) 2015;3(2):1159 [FREE Full text] [doi: 10.13063/2327-9214.1159] [Medline: 26290891]
106. Kheterpal S, Shanks A, Tremper K. Impact of a novel multiparameter decision support system on intraoperative processes of care and postoperative outcomes. Anesthesiology 2018 Feb;128(2):272-282 [FREE Full text] [doi: 10.1097/ALN.0000000000002023] [Medline: 29337743]
107. Thursky KA, Mahemoff M. User-centered design techniques for a computerised antibiotic decision support system in an intensive care unit. Int J Med Inform 2007 Oct;76(10):760-768. [doi: 10.1016/j.ijmedinf.2006.07.011] [Medline: 16950650]
108. Nelson O, Sturgis B, Gilbert K, Henry E, Clegg K, Tan JM, et al. A visual analytics dashboard to summarize serial anesthesia records in pediatric radiation treatment. Appl Clin Inform 2019 Aug 07;10(4):563-569 [FREE Full text] [doi: 10.1055/s-0039-1693712] [Medline: 31390667]
109. Monico LB, Ludwig A, Lertch E, Mitchell SG. Using timeline methodology to visualize treatment trajectories of youth and young adults following inpatient opioid treatment. Int J Qual Methods 2020 Dec;19:160940692097010 [FREE Full text] [doi: 10.1177/1609406920970106]

110. Miller A, Scheinkestel C, Steele C. The effects of clinical information presentation on physicians' and nurses' decision-making in ICUs. Appl Ergon 2009 Jul;40(4):753-761. [doi: 10.1016/j.apergo.2008.07.004] [Medline: 18834970]
111. Hsu CC, Karnwal S, Mullainathan S, Obermeyer Z, Tan C. Characterizing the value of information in medical notes. arXiv. Preprint posted online December 9, 2020 2020 [FREE Full text] [doi: 10.18653/v1/2020.findings-emnlp.187]
112. Arnold CW, Oh A, Chen S, Speier W. Evaluating topic model interpretability from a primary care physician perspective. Comput Methods Programs Biomed 2016 Feb;124:67-75 [FREE Full text] [doi: 10.1016/j.cmpb.2015.10.014] [Medline: 26614020]
113. Hirsch J, Tanenbaum J, Lipsky Gorman S, Liu C, Schmitz E, Hashorva D, et al. HARVEST, a longitudinal patient record summarizer. J Am Med Inform Assoc 2015 Mar;22(2):263-274 [FREE Full text] [doi: 10.1136/amiajnl-2014-002945] [Medline: 25352564]
114. Dudko A, er, Endrjukaite T, Kiyoki Y. Medical documents processing for summary generation and keywords highlighting based on natural language processing and ontology graph descriptor approach. In: Proceedings of the 19th International Conference on Information Integration and Web-based Applications & Services. 2017 Presented at: iiWAS '17; December 4-6, 2017; Salzburg, Austri p. 58-65 URL: https://dl.acm.org/doi/10.1145/3151759.3151784 [doi: 10.1145/3151759.3151784]
115. Baldwin T, Guo Y, Mukherjee VV, Syeda-Mahmood T. Generalized extraction and classification of span-level clinical phrases. AMIA Annu Symp Proc 2018;2018:205-214 [FREE Full text] [Medline: 30815058]
116. Kenei J, Opiyo E, Oboko R. Visualizing semantic structure of a clinical text document. Eur J Electr Eng Comput Sci 2020 Dec 04;4(6) [FREE Full text] [doi: 10.24018/ejece.2020.4.6.256]
117. Dagliati A, Sacchi L, Tibollo V, Cogni G, Teliti M, Martinez-Millana A, et al. A dashboard-based system for supporting diabetes care. J Am Med Inform Assoc 2018 May 01;25(5):538-547 [FREE Full text] [doi: 10.1093/jamia/ocx159] [Medline: 29409033]
118. Carr LL, Zelarney P, Meadows S, Kern JA, Long MB, Kern E. Development of a cancer care summary through the electronic health record. J Oncol Pract 2016 Feb;12(2):e231-e240. [doi: 10.1200/jop.2015.006890]
119. Ham PB, Anderton T, Gallaher R, Hyrman M, Simmerman E, Ramanathan A, et al. Development of electronic medical record-based “rounds report” results in improved resident efficiency, more time for direct patient care and education, and less resident duty hour violations. Am Surg 2016 Sep 01;82(9):853-859 [FREE Full text] [doi: 10.1177/000313481608200950]
120. Klann J, McCoy A, Wright A, Wattanasin N, Sittig D, Murphy S. Health care transformation through collaboration on open-source informatics projects: integrating a medical applications platform, research data repository, and patient summarization. Interact J Med Res 2013 May 30;2(1):e11 [FREE Full text] [doi: 10.2196/ijmr.2454] [Medline: 23722634]
121. Flohr L, Beaudry S, Johnson KT, West N, Burns CM, Ansermino JM, et al. Clinician-driven design of VitalPAD-an intelligent monitoring and communication device to improve patient safety in the intensive care unit. IEEE J Transl Eng Health Med 2018 Mar 05;6:3000114 [FREE Full text] [doi: 10.1109/JTEHM.2018.2812162] [Medline: 29552425]
122. Guo S, Xu K, Zhao R, Gotz D, Zha H, Cao N. EventThread: visual summarization and stage analysis of event sequence data. IEEE Trans Vis Comput Graph 2018 Jan;24(1):56-65. [doi: 10.1109/TVCG.2017.2745320] [Medline: 28866586]
123. Monroe M, Lan R, Lee H, Plaisant C, Shneiderman B. Temporal event sequence simplification. IEEE Trans Vis Comput Graph 2013 Dec;19(12):2227-2236. [doi: 10.1109/TVCG.2013.200] [Medline: 24051789]
124. Klimov D, Shahar Y, Taieb-Maimon M. Intelligent visualization and exploration of time-oriented data of multiple patients. Artif Intell Med 2010 May;49(1):11-31. [doi: 10.1016/j.artmed.2010.02.001] [Medline: 20303245]
125. Brich N, Schulz C, Peter J, Klingert W, Schenk M, Weiskopf D, et al. Visual analytics of multivariate intensive care time series data. Comput Graph Forum 2022 Apr 13;41(6):273-286 [FREE Full text] [doi: 10.1111/cgf.14498]
126. Salvi E, Bosoni P, Tibollo V, Kruijver L, Calcaterra V, Sacchi L, et al. Patient-generated health data integration and advanced analytics for diabetes management: the AID-GM platform. Sensors (Basel) 2019 Dec 24;20(1):128 [FREE Full text] [doi: 10.3390/s20010128] [Medline: 31878195]
127. Drews FA, Doig A. Evaluation of a configural vital signs display for intensive care unit nurses. Hum Factors 2014 May;56(3):569-580. [doi: 10.1177/0018720813499367] [Medline: 24930176]
128. Sacchi L, Capozzi D, Bellazzi R, Larizza C. JTSA: an open source framework for time series abstractions. Comput Methods Programs Biomed 2015 Oct;121(3):175-188. [doi: 10.1016/j.cmpb.2015.05.006] [Medline: 26120073]
129. Hunter J, Freer Y, Gatt A, Logie R, McIntosh N, van der Meulen M, et al. Summarising complex ICU data in natural language. AMIA Annu Symp Proc 2008 Nov 06;2008:323-327 [FREE Full text] [Medline: 18998961]
130. Hunter J, Freer Y, Gatt A, Reiter E, Sripada S, Sykes C. Automatic generation of natural language nursing shift summaries in neonatal intensive care: BT-Nurse. Artif Intell Med 2012 Nov;56(3):157-172. [doi: 10.1016/j.artmed.2012.09.002] [Medline: 23068882]
131. Scott D, Hallett C, Fettiplace R. Data-to-text summarisation of patient records: using computer-generated summaries to access patient histories. Patient Educ Couns 2013 Aug;92(2):153-159 [FREE Full text] [doi: 10.1016/j.pec.2013.04.019] [Medline: 23746770]
132. Shahar Y, Goren-Bar D, Boaz D, Tahan G. Distributed, intelligent, interactive visualization and exploration of time-oriented clinical data and their abstractions. Artif Intell Med 2006 Oct;38(2):115-135. [doi: 10.1016/j.artmed.2005.03.001] [Medline: 16343873]

133. Levy-Fix G, Zucker J, Stojanovic K, Elhadad N. Towards patient record summarization through joint phenotype learning in HIV patients. arXiv. Preprint posted online March 9, 2020 2020 [FREE Full text]
134. Byrne CA, O'Grady M, Collier R, O'Hare GM. An evaluation of graphical formats for the summary of activities of daily living (ADLs). Healthcare (Basel) 2020 Jul 01;8(3):194 [FREE Full text] [doi: 10.3390/healthcare8030194] [Medline: 32630304]
135. Zhang Y, Chanana K, Dunne C. IDMVis: temporal event sequence visualization for type 1 diabetes treatment decision support. IEEE Trans Vis Comput Graph (Forthcoming) 2018 Aug 20. [doi: 10.1109/TVCG.2018.2865076] [Medline: 30136981]
136. Welch G, Balder A, Zagarins S. Telehealth program for type 2 diabetes: usability, satisfaction, and clinical usefulness in an urban community health center. Telemed J E Health 2015 May;21(5):395-403. [doi: 10.1089/tmj.2014.0069] [Medline: 25748544]
137. Sultanum N, Singh D, Brudno M, Chevalier F. Doccurate: a curation-based approach for clinical text visualization. IEEE Trans Vis Comput Graph (Forthcoming) 2018 Aug 20. [doi: 10.1109/TVCG.2018.2864905] [Medline: 30136959]
138. Stubbs B, Kale D, Das A. Sim•TwentyFive: an interactive visualization system for data-driven decision support. AMIA Annu Symp Proc 2012;2012:891-900 [FREE Full text] [Medline: 23304364]
139. Lamy J, Duclos C, Hamek S, Beuscart-Zéphir MC, Kerdelhué G, Darmoni S, et al. Towards iconic language for patient records, drug monographs, guidelines and medical search engines. Stud Health Technol Inform 2010;160(Pt 1):156-160. [Medline: 20841669]
140. Albert R, Agutter J, Syroid N, Johnson K, Loeb R, Westenskow D. A simulation-based evaluation of a graphic cardiovascular display. Anesth Analg 2007 Nov;105(5):1303-1311. [doi: 10.1213/01.ane.0000282823.76059.ca] [Medline: 17959959]
141. Anders S, Albert R, Miller A, Weinger MB, Doig AK, Behrens M, et al. Evaluation of an integrated graphical display to promote acute change detection in ICU patients. Int J Med Inform 2012 Dec;81(12):842-851 [FREE Full text] [doi: 10.1016/j.ijmedinf.2012.04.004] [Medline: 22534099]
142. Effken JA, Loeb RG, Kang Y, Lin ZC. Clinical information displays to improve ICU outcomes. Int J Med Inform 2008 Nov;77(11):765-777. [doi: 10.1016/j.ijmedinf.2008.05.004] [Medline: 18639487]
143. Faiola A, Newlon C. Advancing critical care in the ICU: a human-centered biomedical data visualization systems. In: Proceedings of the 2011 International Conference on Ergonomics and Health Aspects of Work with Computers. 2011 Presented at: EHAWC '11; July 9-14, 2011; Orlando, FL p. 119-128 URL: https://link.springer.com/chapter/10.1007/978-3-642-21716-6_13 [doi: 10.1007/978-3-642-21716-6_13]
144. Faiola A, Srinivas P, Duke J. Supporting clinical cognition: a human-centered approach to a novel ICU information visualization dashboard. AMIA Annu Symp Proc 2015;2015:560-569 [FREE Full text] [Medline: 26958190]
145. Forsman J, Anani N, Eghdam A, Falkenhav M, Koch S. Integrated information visualization to support decision making for use of antibiotics in intensive care: design and usability evaluation. Inform Health Soc Care 2013 Dec;38(4):330-353. [doi: 10.3109/17538157.2013.812649] [Medline: 23957739]
146. Görges M, Kück K, Koch SH, Agutter J, Westenskow DR. A far-view intensive care unit monitoring display enables faster triage. Dimens Crit Care Nurs 2011;30(4):206-217. [doi: 10.1097/DCC.0b013e31821b7f08] [Medline: 21654229]
147. Wachter SB, Johnson K, Albert R, Syroid N, Drews F, Westenskow D. The evaluation of a pulmonary display to detect adverse respiratory events using high resolution human simulator. J Am Med Inform Assoc 2006 Nov;13(6):635-642 [FREE Full text] [doi: 10.1197/jamia.M2123] [Medline: 16929038]
148. van Amsterdam K, Cnossen F, Ballast A, Struys M. Visual metaphors on anaesthesia monitors do not improve anaesthetists' performance in the operating theatre. Br J Anaesth 2013 May;110(5):816-822 [FREE Full text] [doi: 10.1093/bja/aes516] [Medline: 23384736]
149. Ordonez P, Oates T, Lombardi ME, Hernandez G, Holmes K, Fackler J, et al. Visualization of multivariate time-series data in a neonatal ICU. IBM J Res Dev 2012 Sep;56(5):7:1-712 [FREE Full text] [doi: 10.1147/JRD.2012.2200431]
150. McKinlay A, McVittie C, Reiter E, Freer Y, Sykes C, Logie R. Design issues for socially intelligent user interfaces. A discourse analysis of a data-to-text system for summarizing clinical data. Methods Inf Med 2010;49(4):379-387. [doi: 10.3414/ME0613] [Medline: 20027380]
151. Jadhav A, Baldwin T, Wu J, Mukherjee V, Syeda-Mahmood T. Semantic expansion of clinician generated data preferences for automatic patient data summarization. AMIA Annu Symp Proc 2021;2021:571-580 [FREE Full text] [Medline: 35308964]
152. Laxmisan A, McCoy A, Wright A, Sittig D. Clinical summarization capabilities of commercially-available and internally-developed electronic health records. Appl Clin Inform 2017 Dec 16;03(01):80-93 [FREE Full text] [doi: 10.1055/s-0037-1618556]
153. Lin C. ROUGE: a package for automatic evaluation of summaries. Text Summarization Branches Out. URL: https://aclanthology.org/W04-1013/ [accessed 2022-11-15]
154. Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. 2002 Presented at: ACL '02; July 7-12, 2002; Philadelphia, PA p. 311-318 URL: https://dl.acm.org/doi/10.3115/1073083.1073135 [doi: 10.3115/1073083.1073135]
155. Rimmer A. Radiologist shortage leaves patient care at risk, warns royal college. BMJ 2017 Oct 11;359:j4683. [doi: 10.1136/bmj.j4683] [Medline: 29021184]

156. Zhu X, Cimino JJ. Clinicians' evaluation of computer-assisted medication summarization of electronic medical records. Comput Biol Med 2015 Apr;59:221-231 [FREE Full text] [doi: 10.1016/j.compbiomed.2013.12.006] [Medline: 24393492]
157. Pivovarov R, Coppleson YJ, Gorman SL, Vawdrey DK, Elhadad N. Can patient record summarization support quality metric abstraction? AMIA Annu Symp Proc 2016;2016:1020-1029 [FREE Full text] [Medline: 28269899]
158. Van Vleck TT, Stein DM, Stetson PD, Johnson SB. Assessing data relevance for automated generation of a clinical summary. AMIA Annu Symp Proc 2007 Oct 11;2007:761-765 [FREE Full text] [Medline: 18693939]
159. El-Kassas WS, Salama CR, Rafea AA, Mohamed HK. Automatic text summarization: a comprehensive survey. Expert Syst Appl 2021 Mar;165:113679. [doi: 10.1016/j.eswa.2020.113679]
160. Tan J, Wan X, Xiao J. From neural sentence summarization to headline generation: a coarse-to-fine approach. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence. 2017 Presented at: IJCAI'17; August 19-25, 2017; Melbourne, Australia p. 4109-4115 URL: https://dl.acm.org/doi/abs/10.5555/3171837.3171860 [doi: 10.24963/ijcai.2017/574]
161. Hou L, Hu P, Bei C. Abstractive document summarization via neural model with joint attention. In: Proceedings of the 2017 National CCF conference on natural language processing and Chinese computing. 2017 Presented at: NLPCC '17; November 8-12, 2017; Dalian, China p. 329-338 URL: https://link.springer.com/chapter/10.1007/978-3-319-73618-1_28 [doi: 10.1007/978-3-319-73618-1_28]
162. Moratanch N, Chitrakala S. A survey on extractive text summarization. In: Proceedings of the 2017 International Conference on Computer, Communication and Signal Processing. 2017 Presented at: ICCCSP '17; January 10-11, 2017; Chennai, India p. 1-6 URL: https://ieeexplore.ieee.org/document/7944061 [doi: 10.1109/icccsp.2017.7944061]
163. Gupta V, Lehal GS. A survey of text summarization extractive techniques. J Emerg Technol Web Intell 2010 Aug 20;2(3):258-268 [FREE Full text] [doi: 10.4304/jetwi.2.3.258-268]
164. Savova G, Masanz J, Ogren P, Zheng J, Sohn S, Kipper-Schuler K, et al. Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc 2010;17(5):507-513 [FREE Full text] [doi: 10.1136/jamia.2009.001560] [Medline: 20819853]
165. Alipour S, Eslami B, Abedi M, Ahmadinejad N, Arabkheradmand A, Aryan A, et al. A practical, clinical user-friendly format for breast ultrasound report. Eur J Breast Health 2021 Apr 1;17(2):165-172 [FREE Full text] [doi: 10.4274/ejbh.galenos.2021.6344] [Medline: 33870117]
166. de Baca ME, Arnaout R, Brodsky V, Birdsong GG. Ordo ab Chao: framework for an integrated disease report. Arch Pathol Lab Med 2015 Feb;139(2):165-170 [FREE Full text] [doi: 10.5858/arpa.2013-0561-CP] [Medline: 25611099]
167. Kay S. The international patient summary and the summarization requirement. Stud Health Technol Inform 2021 Oct 27;285:17-30. [doi: 10.3233/SHTI210569] [Medline: 34734848]
168. Charette RS, Sarpong NO, Weiner TR, Shah RP, Cooper HJ. What's in a summary? Laying the groundwork for advances in hospital-course summarization. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021 Presented at: NAACL '21; June 6–11, 2021; Virtual Event p. 4794-4811 URL: https://aclanthology.org/2021.naacl-main.382.pdf [doi: 10.52198/22.sti.41.os1633]
169. Law AS, Freer Y, Hunter J, Logie RH, McIntosh N, Quinn J. A comparison of graphical and textual presentations of time series data to support medical decision making in the neonatal intensive care unit. J Clin Monit Comput 2005 Jun;19(3):183-194. [doi: 10.1007/s10877-005-0879-3] [Medline: 16244840]
170. Bauer DT, Guerlain S, Brown PJ. The design and evaluation of a graphical display for laboratory data. J Am Med Inform Assoc 2010 Jul 01;17(4):416-424 [FREE Full text] [doi: 10.1136/jamia.2009.000505] [Medline: 20595309]
171. Salmon P, Rappaport A, Bainbridge M, Hayes G, Williams J. Taking the problem oriented medical record forward. Proc AMIA Annu Fall Symp 1996:463-467 [FREE Full text] [Medline: 8947709]
172. Zhou L, Hripcsak G. Temporal reasoning with medical data--a review with emphasis on medical natural language processing. J Biomed Inform 2007 Apr;40(2):183-202 [FREE Full text] [doi: 10.1016/j.jbi.2006.12.009] [Medline: 17317332]
173. Dziadzko MA, Herasevich V, Sen A, Pickering BW, Knight AA, Moreno Franco P. User perception and experience of the introduction of a novel critical care patient viewer in the ICU setting. Int J Med Inform 2016 Apr;88:86-91. [doi: 10.1016/j.ijmedinf.2016.01.011] [Medline: 26878767]
174. Li W, McCallum A. Pachinko allocation: DAG-structured mixture models of topic correlations. In: Proceedings of the 23rd international conference on Machine learning. 2006 Presented at: ICML '06; June 25-29, 2006; Pittsburgh, PA p. 577-584 URL: https://dl.acm.org/doi/10.1145/1143844.1143917 [doi: 10.1145/1143844.1143917]
175. Plaisant C, Mushlin R, Snyder A, Li J, Heller D, Shneiderman B. LifeLines: using visualization to enhance navigation and analysis of patient records. Craft Inf Vis 2003:308-312 [FREE Full text] [doi: 10.1016/b978-155860915-0/50038-x]
176. Goldstein A, Shahar Y. Generation of natural-language textual summaries from longitudinal clinical records. In: Sarkar IN, Georgiou A, de Azevedo Marques PM, editors. Studies in Health Technology and Informatics. Amsterdam, Netherlands: ISO Press; 2015:594.
177. Ghosh A. On the challenges of using evidence-based information: the role of clinical uncertainty. J Lab Clin Med 2004 Aug;144(2):60-64. [doi: 10.1016/j.lab.2004.05.013] [Medline: 15322499]
178. Chang H. Evaluation framework for telemedicine using the logical framework approach and a fishbone diagram. Healthc Inform Res 2015 Oct;21(4):230-238 [FREE Full text] [doi: 10.4258/hir.2015.21.4.230] [Medline: 26618028]

179. Lin YL, Guerguerian A, Laussen P. Heuristic evaluation of data integration and visualization software used for continuous monitoring to support intensive care: a bedside nurse's perspective. J Nurs Care 2015;04(06):1-8 [FREE Full text] [doi: 10.4172/2167-1168.1000300]
180. Hsueh PS, Zhu XX, Hsiao MJ, Lee SY, Deng V, Ramakrishnan S. Automatic summarization of risk factors preceding disease progression an insight-driven healthcare service case study on using medical records of diabetic patients. World Wide Web 2014 Sep 14;18(4):1163-1175 [FREE Full text] [doi: 10.1007/s11280-014-0304-2]
181. Chaudhary A, George M, Chacko A. Extractive summarization of EHR notes. In: Proceedings of the 2020 International Conference on Paradigms of Computing, Communication and Data Sciences. 2020 Presented at: PCCDS '20; May 1-3, 2020; Kurukshetra, India p. 909-919 URL: https://link.springer.com/chapter/10.1007/978-981-15-7533-4_73 [doi: 10.1007/978-981-15-7533-4_73]
182. Anokwa Y, Ribeka N, Parikh T, Borriello G, Were M. Design of a phone-based clinical decision support system for resource-limited settings. In: Proceedings of the 5th International Conference on Information and Communication Technologies and Development. 2012 Presented at: ICTD '12; March 12-15, 2012; Atlanta, GA p. 13-24 URL: https://dl.acm.org/doi/10.1145/2160673.2160676 [doi: 10.1145/2160673.2160676]

Abbreviations

EHR: electronic health record
HCP: health care professional
ICU: intensive care unit
LIME: Local Interpretable Model-Agnostic Explanations
NLP: natural language processing
PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
ROUGE: Recall-orientated Understudy for Gisting Evaluation
UMLS: Unified Medical Language System

Edited by A Benis; submitted 28.11.22; peer-reviewed by R Perotte, PF Chen, E Sezgin, X Yan; comments to author 26.12.22; revised version received 15.03.23; accepted 25.07.23; published 28.11.23.

Please cite as:
Keszthelyi D, Gaudet-Blavignac C, Bjelogrlic M, Lovis C
Patient Information Summarization in Clinical Settings: Scoping Review
JMIR Med Inform 2023;11:e44639
URL: https://medinform.jmir.org/2023/1/e44639
doi:10.2196/44639
PMID:38015588

Review

Applying Natural Language Processing to Textual Data From Clinical Data Warehouses: Systematic Review

Adrien Bazoge[1,2], MSc; Emmanuel Morin[1], PhD; Béatrice Daille[1], PhD; Pierre-Antoine Gourraud[2,3], PhD

[1]Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France

[2]Nantes Université, CHU de Nantes, Pôle Hospitalo-Universitaire 11: Santé Publique, Clinique des données, INSERM, CIC 1413, F-44000 Nantes, France

[3]Nantes Université, INSERM, CHU de Nantes, École Centrale Nantes, Centre de Recherche Translationnelle en Transplantation et Immunologie, CR2TI, F-44000 Nantes, France

Corresponding Author:
Pierre-Antoine Gourraud, PhD
Nantes Université, INSERM, CHU de Nantes, École Centrale Nantes, Centre de Recherche Translationnelle en Transplantation et Immunologie, CR2TI
30 bvd Jean Monnet - 2eme étage
F-44000 Nantes
France
Phone: 33 2 447 68 234
Email: Pierre-Antoine.Gourraud@univ-nantes.fr

Abstract

Background: In recent years, health data collected during the clinical care process have been often repurposed for secondary use through clinical data warehouses (CDWs), which interconnect disparate data from different sources. A large amount of information of high clinical value is stored in unstructured text format. Natural language processing (NLP), which implements algorithms that can operate on massive unstructured textual data, has the potential to structure the data and make clinical information more accessible.

Objective: The aim of this review was to provide an overview of studies applying NLP to textual data from CDWs. It focuses on identifying the (1) NLP tasks applied to data from CDWs and (2) NLP methods used to tackle these tasks.

Methods: This review was performed according to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines. We searched for relevant articles in 3 bibliographic databases: PubMed, Google Scholar, and ACL Anthology. We reviewed the titles and abstracts and included articles according to the following inclusion criteria: (1) focus on NLP applied to textual data from CDWs, (2) articles published between 1995 and 2021, and (3) written in English.

Results: We identified 1353 articles, of which 194 (14.34%) met the inclusion criteria. Among all identified NLP tasks in the included papers, information extraction from clinical text (112/194, 57.7%) and the identification of patients (51/194, 26.3%) were the most frequent tasks. To address the various tasks, symbolic methods were the most common NLP methods (124/232, 53.4%), showing that some tasks can be partially achieved with classical NLP techniques, such as regular expressions or pattern matching that exploit specialized lexica, such as drug lists and terminologies. Machine learning (70/232, 30.2%) and deep learning (38/232, 16.4%) have been increasingly used in recent years, including the most recent approaches based on transformers. NLP methods were mostly applied to English language data (153/194, 78.9%).

Conclusions: CDWs are central to the secondary use of clinical texts for research purposes. Although the use of NLP on data from CDWs is growing, there remain challenges in this field, especially with regard to languages other than English. Clinical NLP is an effective strategy for accessing, extracting, and transforming data from CDWs. Information retrieved with NLP can assist in clinical research and have an impact on clinical practice.

(JMIR Med Inform 2023;11:e42477) doi:10.2196/42477

KEYWORDS

natural language processing; data warehousing; clinical data warehouse; artificial intelligence; AI

Introduction

Background

For >20 years, health data from patient care have been systematically archived in the form of electronic health records (EHRs) [1,2]. Databases have been created to gather both structured data (eg, vital signs and clinical-biological characteristics and demographics) and unstructured data (eg, textual reports of hospitalizations or visits). These large amounts of data involve multiple contributors: patients, for whom data are collected during hospitalizations or visits; caregivers, who care for the patients and collect the data; and health care institutions, which organize all operational and financial logistics involving the care and related data [3]. The first purpose of collecting these data is to broadly deliver high-quality care to patients, even if the data may be repurposed for secondary use, such as reduction in health care costs, population health management, and clinical research [1]. Human data in clinical research are intended for research purposes and limited in terms of sample size, scope, and longitudinal follow-up (ie, clinical trials or disease registries). The secondary use of EHRs allows to increase patient recruitment in trials [4] and enables access to a larger variety of clinical information for clinical research [5,6].

The rapid increase in digital data production prompted the construction of clinical data warehouses (CDWs), also known as health data warehouses or biomedical data warehouses, for the secondary use of EHRs [2]. *CDW* refers to the interconnection of disparate data from different sources, which are restructured into a common format and indexed using standard vocabularies. CDWs collect data from millions of patients treated in hospitals and can be accessed by stakeholders to analyze care situations and make critical decisions [7]. Unlike in the fields of logistics, marketing, and sales, the health care field has been slow to fully integrate data warehouses. CDWs require managing security and privacy constraints related to medical data [7]. Depending on which country houses the CDW, medical data–related policies can vary and potentially slow the construction process [8]. Data warehouses have been part of the health care landscape for decades [9], especially in the United States, where the first CDWs appeared in the 1990s. In some countries, such as France, CDWs have only been constructed more recently owing to policy constraints. At the institutional level, the use of CDWs underscores that organizations recognize the transformative potential and value of the data generated by their activity. This secondary use of data is facilitated by technological advances in artificial intelligence [10]. Among many types of data, textual data reinforce the popularity of a subgroup of artificial intelligence methods, natural language processing (NLP), which implements algorithms that can operate on massive unstructured textual data [11]. The majority of clinical information is stored in unstructured text format, and NLP allows accessing this information [12,13].

Objectives

This review aims at providing an overview of studies applying clinical NLP to textual data from CDWs. The focus of this review is to identify the (1) NLP tasks applied to data from CDWs and (2) NLP methods used for each task.

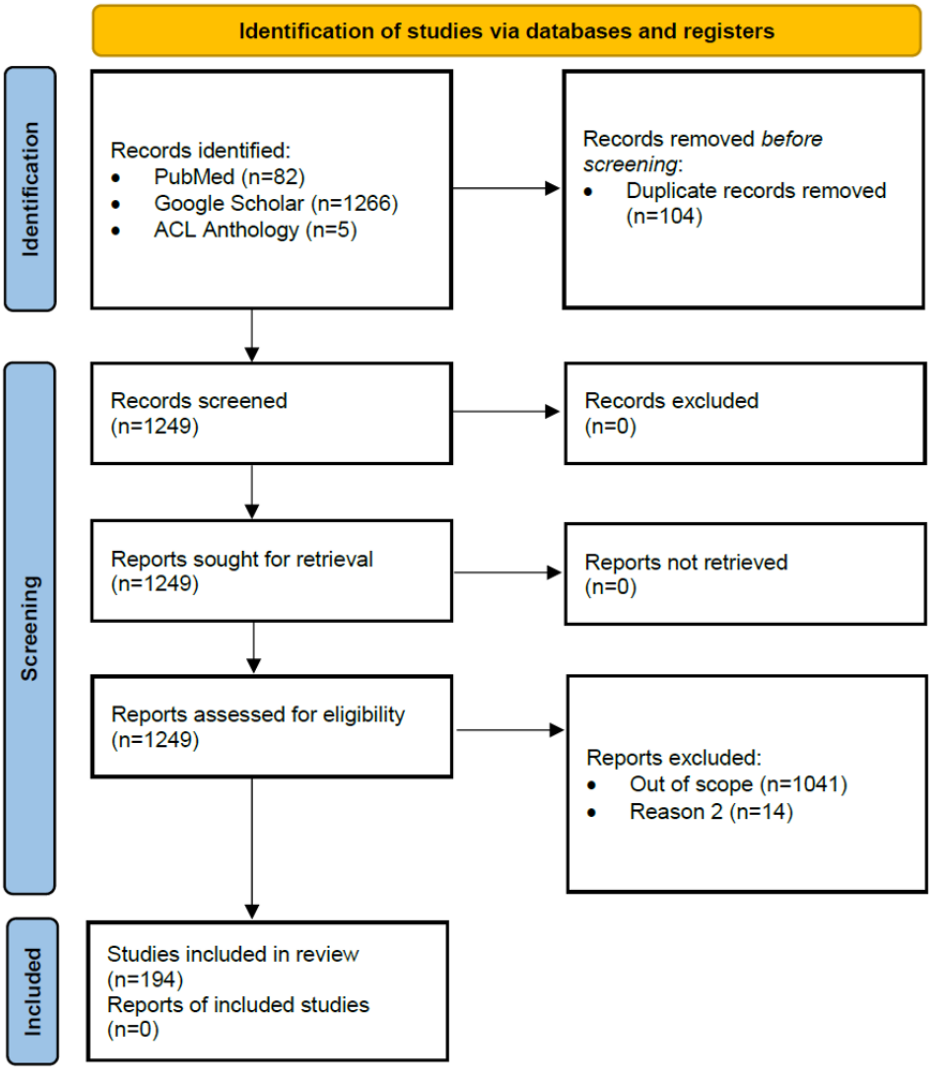
Methods

The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines were followed for reporting this review (Multimedia Appendix 1).

Review Method and Selection Criteria

Articles identified from the queries were manually included on the basis of the following inclusion criteria: articles (1) mentioning the use of NLP on data from CDWs, (2) published between 1995 and 2021, and (3) written in English. The inclusion was carried out by reading titles and abstracts or by searching the article for the keywords used in the queries to determine whether it was relevant. Details of the article selection steps are described in Figure 1.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) article selection flowchart.



Bibliographic Databases

We searched for relevant articles in 3 bibliographic databases: PubMed, ACL Anthology, and Google Scholar. PubMed is specialized in biomedical literature; its query builder allows searchers to construct queries based on both Medical Subject Headings terms and natural language. ACL Anthology covers the literature published in conferences related to computational linguistics and NLP. Google Scholar does not have a dedicated area of specialty for the papers it references and covers a wide range of the literature.

Search Strategy

Identifying papers with NLP applied to data from CDWs involved combining multiple designations: the term *data warehouse* is sometimes referred to as a *database* or a *repository*. In addition, the source of the data used in clinical studies may only be listed in the main manuscript. Data collection requires using multiple queries to aim at both high specificity and high sensitivity.

To retrieve a representative selection of papers, we used queries based on specific keywords for each topic of interest, that is, (1) CDWs and (2) NLP:

- CDWs: “clinical data warehouse,” “biomedical data warehouse,” and “health data warehouse.” The selected keywords representing this topic correspond to the most commonly used terms for CDWs.
- NLP: “natural language processing,” “NLP,” and “text mining.” The keyword “text mining” complements the concept of the “natural language processing” keyword. Text mining stands out as the most frequently used NLP application in the medical field. As a result, the term “natural language processing” can sometimes be eclipsed by “text mining.”

Several queries were made using the selected keywords in each bibliographic database. The details of each query are available in Multimedia Appendix 2.

All queries were run on February 23, 2022. PubMed and ACL Anthology papers were retrieved by manually executing queries on the respective websites of these bibliographic databases. Google Scholar papers were collected using free software [14]. The results of the queries were merged, and duplicates were removed.

The queries are not exhaustive but rather aim to provide a limited and representative selection of papers covering the topics

of interest. Synonyms for *warehouse*, such as *database* or *repository*, were not used in the queries to avoid the collection of a significant number of irrelevant articles to review. Furthermore, some papers may also apply NLP to data from CDWs without mentioning the CDW and could be missed by the queries.

Data Collection

The following data were manually collected from the included articles: (1) NLP tasks addressed in the original paper (the NLP task classification is based on the one provided by Névéol et al [13]), (2) NLP methods used to address the tasks, (3) the CDW that is the source of the data, and (4) the language of the data used in the paper.

Results

Overview

A total of 1353 articles (PubMed: n=82, 6.06%; Google Scholar: n=1266, 93.57%; and ACL Anthology: n=5, 0.37%) were identified with the initial search strategy. After reviewing the title and abstract of each article, of the 1353 articles, 1159 (85.66%) were excluded owing to duplication (n=104, 8.97%), language issues (n=14, 1.21%), and for being out of the scope of this review (n=1041, 89.82%). Overall, of the initially identified 1353 articles, 194 (14.34%) met the inclusion criteria. These 194 articles were published between 2002 and 2021, which means that articles published between 1995 and 2001 did not meet the inclusion criteria.

This section gathers the topics covered in published research on NLP applied to data from CDWs. The results of the reviewed articles are presented by the NLP task mentioned in the articles. Although many articles address the same NLP task, we decided to not directly compare the performances of the methods used in the articles in this review. Methods have been evaluated with different data in different languages and with different metrics. Hence, we concluded that it was not relevant to perform this comparison.

Table 1 gives the count of studies based on the NLP task for 2 periods of time: 2002-2015 and 2016-2021. The 2 time periods were chosen owing to the transition in the NLP paradigm, shifting from knowledge-based to machine learning methods. This transition coincided with the emergence of new tasks, including language modeling.

Table 1. Natural language processing (NLP) tasks reported in the retrieved publications (n=194).

NLP tasks	NLP methods used, n (%)		References
	2002-2015	2016-2021	
Information extraction (n=112)			
Medical concepts (n=37)	S[a]: 14 (74); ML[b]: 5 (26)	S: 10 (40); ML: 11 (44); DL[c]: 4 (16)	[15-51]
Specific characteristics (n=40)	S: 4 (67); ML: 2 (33)	S: 22 (56); ML: 12 (31); DL: 5 (13)	[52-91]
Drugs and adverse events (n=26)	S: 10 (77); ML: 3 (23)	S: 8 (57); ML: 1 (7); DL: 5 (36)	[49,52,92-115]
Findings and symptoms (n=8)	S: 1 (50); ML: 1 (50)	S: 2 (25); ML: 2 (25); DL: 4 (50)	[49,52,116-121]
Relation extraction (n=1)	S: 1 (100)	N/A[d]	[50]
Classification (n=51)			
Phenotyping (n=38)	S: 7 (78); ML: 2 (22)	S: 17 (49); ML: 12 (34); DL: 6 (17)	[50,122-158]
Indexing and coding (n=7)	S: 3 (100)	S: 2 (50); ML: 1 (25); DL: 1 (25)	[159-165]
Topic modeling (n=3)	N/A	S: 1 (25); ML: 3 (75)	[166-168]
Patient identification (n=3)	N/A	S: 1 (25); ML: 2 (50); DL: 1 (25)	[169-171]
Context analysis (n=18)			
Similarity (n=6)	S: 2 (100)	S: 1 (25); DL: 3 (75)	[172-177]
Temporality (n=4)	S: 1 (100)	S: 2 (100)	[93,178-180]
Negation detection (n=3)	N/A	S: 2 (67); DL: 1 (33)	[178,181,182]
Abbreviation (n=2)	N/A	S: 2 (100)	[183,184]
Uncertainty (n=1)	N/A	S: 1 (100)	[180]
Experiencer (n=2)	N/A	S: 2 (100)	[178,182]
Language modeling (n=11)	N/A	ML: 6 (46); DL: 7 (54)	[171,185-194]
Resource development (n=6)			
Corpora and annotation (n=4)	N/A	ML: 1 (100)	[195-198]
Lexica (n=2)	N/A	S: 2 (67); ML: 1 (33)	[199,200]
Shared tasks (n=5)	S: 4 (57); ML: 3 (43)	S: 1 (100)	[201-205]
Deidentification (n=2)	S: 1 (50); ML: 1 (50)	DL: 1 (100)	[206,207]
Data cleaning (n=1)	N/A	ML: 1 (100)	[208]

[a]S: symbolic methods.

[b]ML: machine learning.

[c]DL: deep learning.

[d]N/A: not applicable.

Information Extraction

Information extraction is one of the most studied tasks in NLP within the clinical field. In the included articles, named entity recognition (NER) primarily focuses on identifying entities such as protected health information (PHI) to deidentify clinical documents [206,207], as well as various clinical concepts. These concepts encompass diseases [20,25,40,41,45,47,49]; findings and symptoms [49,52,116-119,121]; and medication names [49,52,93-95,99,100,102,106,107,112,113,115], along with their associated details such as dose, frequency, and duration [52,93-95,112,113,115] as well as potential adverse events [96-98,100,101,106-110,114]. These medical concepts can be mapped to terminologies or ontologies such as the Unified Medical Language System (UMLS) [23,24,30,37-39,41,46,97], Systematized Nomenclature of Medicine–Clinical Terms (SNOMED-CT) [27,28,30], or International Classification of Diseases, Ninth Revision (ICD-9) [21].

Several popular NLP systems have been extensively used for extracting, structuring, and encoding clinical information from narrative patient reports in English. Numerous studies detail the application of the Medical Language Extraction and Encoding System (MedLEE) for clinical concepts [24,27-29,32-36,50,51,121] or medication [103,104,111] extraction, as well as UMLS coding. The extraction and mapping of clinical information from clinical notes to UMLS has also been accomplished using the clinical Text Analysis and Knowledge Extraction System (cTAKES) [16,17,20,22,100,129,134,168], MetaMap [31,37,38,47], MedTagger [44,45,67,78,86,105], and the National Center for Biomedical Ontology (NCBO) Annotator

[97,99,106,107,109,114]. Extracted concepts can be mapped to other standard ontologies and terminologies, such as SNOMED-CT [27]. Caliskan et al [95] evaluated the Averbis Health Discovery NLP system on a medication extraction task on German clinical notes.

Other systems addressing NER or information extraction were customized to specific use cases. Rule-based methods encoded dictionaries and terminologies to match terms and concepts in clinical texts [40-42,49,102,108,112,113]. Machine learning methods take advantage of the clinical knowledge in the large amount of data in CDWs. According to the time period, methods that were used reflect the trend of using NLP state-of-the-art methods and language models. Conditional random fields (CRFs) were used to extract clinical concepts [23,46] or PHI for the deidentification of clinical documents [207]. Hierarchically supervised latent Dirichlet allocation was applied to hospital discharge summaries to predict ICD-9 codes [21]. Deep learning approaches such as bidirectional long short-term memory–CRF (BiLSTM-CRF) [93,113,115] and recurrent neural network grammars [93] performed medical entity extraction in French clinical texts. Chokshi et al [119] compared a bag-of-words model with support vector machine (SVM) and 2 neural network models: a convolutional neural network (CNN) and a neural attention model, both with Word2Vec embedding as input. The accuracies of the CNN and neural attention model models were relatively equal, but they were higher than the accuracy of the SVM model. Lerner et al [49] compared 3 systems for clinical NER: a terminology-based system built on UMLS and SNOMED-CT, a bidirectional gated recurrent unit–CRF system, and a hybrid system using the prediction of the terminology-based system as a feature for the bidirectional gated recurrent unit–CRF system. Yang et al [206] identified PHI from free text with a long short-term memory (LSTM)–CRF model.

Recent state-of-the-art models based on transformer neural architectures [209] were also applied to extract medical concepts. Neuraz et al [52] used a BiLSTM-CRF layer on top of a vector representation of tokens computed by Bidirectional Encoder Representations from Transformers (BERT) in French. BERT and Robustly Optimized BERT Pretraining Approach were examined to extract social and behavioral determinants of health concepts from clinical narratives [15]. Some of the studies paired a neural language model with simple pattern matching techniques; for example, Jouffroy et al [115] proposed a hybrid approach for the extraction of medication information from French clinical text that combined regular expressions to preannotate the text with contextual word embeddings (embeddings from language models [ELMo]) that are fed into a deep recurrent neural network (BiLSTM-CRF).

Some of the studies (31/194, 16%) addressed specific clinical information extracted from clinical texts. These included bone density [59], breast cancer gene 1 or 2 mentions [86], the predictors and timing of lifestyle modification for patients with hypertension [60], the determination of positivity at imaging presentation in radiology reports [66], Banff classification [69], surgical site infection [70], Breast Imaging Reporting and Database System category 3 [71,72], chemotherapy toxicities [76], vital signs [79], transurethral resection of bladder tumors [80], statin use [57], human leukocyte antigen genotypes [82], unplanned episodes of care [83], smoking status [65,84], monoclonal gammopathy [90], skeletal site-specific fractures [85], and social determinants of health [66]. Methods used to extract this information were rule based [67,69-72,76,79,80,82-85], statistical [59,60], or a combination of both [86,90].

Multiple pieces of information about patients were extracted from clinical texts for application in retrospective studies [56]. Ansoborlo et al [89] extracted 52 pieces of bioclinical information from French multidisciplinary team meeting reports concerning lung cancer by applying regular expressions and then compared this approach with a Bayesian classifier method.

Extracting information from clinical text was also carried out as a prediction task. Predicted data cover length of hospital stay [73], the likelihood of neuroscience intensive care unit admission [64], the risk of 30-day readmission in patients with heart failure [55], or quality metrics for the assessment of pretreatment digital rectal examination documentation [62]. Risk assessments of diseases or pathologies, including HIV [61,81], pancreatic cancer [75], pressure ulcer [91], chronic kidney disease [63], and breast cancer [54], have also been studied as prediction tasks. Predicting this clinical information can be achieved with rule-based methods [73,81], machine learning techniques such as latent Dirichlet allocation [63,73], or a combination of both [75,91].

Context Analysis

Linguistic occurrences are particularly relevant where medical information is concerned, such as negation, temporality, uncertainty, or experiencer (ie, determine whether the identified information is related to the patient or a third party, such as a family member). In the included studies, rule-based methods were often used to detect contextual information in clinical text [178,180,182]. Although these methods offer good results (with an approximate F_1-measure value of 0.90), they rely on handmade resources, such as terminologies and regular expressions, and customization is often needed for specific use cases. Temporality patterns have been studied by Liu et al [92] to discern adverse drug events from indications in clinical text. Zhou et al [179] describe a temporal constraint structure constructed from temporal expressions in discharge summaries to model these expressions. In the clinical domain, many temporal expressions have unique characteristics, and this structure provides comprehensive coverage in encoding these expressions. Abbreviations are widely used in medicine and have been studied in French [183] and English [184] clinical texts to better handle medical abbreviations. Recent embedding-based methods such as BERT have made it easier to study negation detection [181] and text similarity [173,174]. Text similarity has also been studied to identify semantically similar concepts [175], similar patients [177], or to detect redundancy in clinical texts [172,176].

Classification

Identifying patients is a key component in clinical research for constructing population studies. NLP can improve the querying and indexing of patients and their data in CDWs. Zhu et al [161]

addressed query expansion based on a large in-domain clinical corpus to solve problems of polysemy, synonymy, and hyponymy in clinical text to improve patient identification. Query expansion was also studied through 3 methods: synonym expansion strategy, topic modeling, and a predicate-based strategy derived from MEDLINE abstracts [165]. An automated electronic search algorithm for identifying postoperative complications was evaluated by Tien et al [162]. A semantic health data warehouse was designed to assist health professionals in prescreening eligible patients in clinical trials [163,164]. A combination of structured and unstructured German data was used by Scheurwegs et al [160] to assign clinical codes to patient stays.

Downstream of the query of CDWs, NLP can be applied to identify patients or documents of interest when the classification methods offered by CDWs are not precise enough. Patient identification can be carried out using methods such as rule-based approaches, which involve using terms related to eligible criteria [127,137,140-150,153,170], or learning-based approaches [126,131,133], or a combination of both [152,155-157,169]. Li et al [166] and Chen et al [167] applied latent Dirichlet allocation in clinical notes for topic modeling. Agarwal et al [154] detailed a logistic regression model of phenotypes learned on noisy labeled data. Some of the studies (4/194, 2.1%) relied on Dr Warehouse, a biomedical data warehouse oriented toward clinical narrative reports, developed at Necker Children's Hospital in Paris, France. This data warehouse was used to explore, using the frequency and term frequency–inverse document frequency (TF-IDF), the association between clinical phenotypes and rare diseases such as the potassium voltage-gated channel subfamily A member 2 variant in neurodevelopmental syndromes [138], Dravet syndrome [125], ciliopathy [139], and other rare diseases [136].

Language Modeling

Recent word embedding–based methods take advantage of the large amount of data stored in CDWs to learn effective semantic representations of clinical texts. In the included articles, these methods allowed to make calculations on words to find, for example, similar terms in the embedding space [88,130]. Among these methods, transformer-based models, such as BERT, were fine-tuned for multiple tasks, including text classification to map document titles to Logical Observation Identifiers Names and Codes Document Ontology [159] and sequence labeling to detect and estimate the location of abnormalities in whole-body scans [53]. Similarly, clinical text was structured with the classification of ICD-9 codes based on vectorization methods [190,191].

Some of the studies evaluated the effectiveness of word embedding models on multiple tasks. Lee et al [135] evaluated Node2Vec, singular value decomposition, Language Identification for Named Entities, Word2Vec, and global vectors for word representation (GloVe) in retrieving relevant medical features for phenotyping tasks. The authors demonstrated that GloVe, when trained on EHR data, outperforms other embedding methods. GloVe and Word2Vec were used in conjunction with LSTM and gated recurrent unit and evaluated across multiple tasks, with gated recurrent unit outperforming LSTM [192]. Similarly, Dynomant et al [193] compared on multiple tasks 3 embedding methods (Word2Vec, GloVe, and fastText) trained on a French corpus. The 3 methods were evaluated on 4 tasks, and Word2Vec with the skip-gram architecture had the highest score on 3 (75%) of the 4 tasks. Peng et al [185] evaluated 2 transformer-based models, BERT and ELMo, on 10 benchmark data sets and found that the BERT model achieved the best results. BERT was also evaluated on sentence similarity, relation extraction, inference, and NER tasks on data sets from clinical domains [186]. The study by Neuraz et al [188] comparing fastText and ELMo showed that models learned on clinical data performed better than models learned on data from the general domain. The study by Tawfik and Spruit [187] described a toolkit to evaluate the effectiveness of sentence representation learning models.

Text representation models are commonly used as embedding layers in neural network models developed for specific tasks. Word2Vec has been used in numerous studies for various purposes, including assessing bone scan use among patients with prostate cancer with a CNN [151], screening and diagnosing of breast cancer with a deep learning architecture [123], extracting features used for risk prediction of liver transplantation for hepatocellular cancer with a capsule neural network [124], and using a CNN to learn the clinical trial criteria eligibility status of patients for participation in cohort studies [171]. Lee et al [194] proposed a unified graph representation learning framework based on graph convolutional networks and LSTM to construct an EHR graph representation of medical entities. Dligach et al [189] developed a clinical text encoder for specific phenotypes. Experiments were conducted with a deep averaging network and a CNN to construct this text encoder.

Resource Development and Shared Tasks

Many NLP methods rely on clinically specific resources to be developed. In the included articles, data from CDWs, combined with clinical expert knowledge, allowed the development of resources such as annotation guidelines and schemes [195,196,198], lexica [200], ontologies [199], or frameworks to validate the outputs of NLP systems [197].

International community efforts have been demonstrated through shared tasks involving clinical notes from CDWs. In the included articles, the Informatics for Integrating Biology and the Bedside (i2b2) obesity challenge focused on obesity and its 15 most common comorbidities through a multiclass multilabel classification task [204,205]. Another i2b2 challenge held in 2009 concerned extracting medication information from clinical text [202,210]. Three tasks were proposed in the fourth i2b2 or Department of Veterans Affairs shared-task and workshop challenge: extraction of medical problems, tests, and treatments; classification of assertions made on medical problems; and classification of a relationship between a pair of concepts that appear in the same sentence where at least 1 concept is a medical problem [202]. These i2b2 shared tasks relied on deidentified discharge summaries from the Partners HealthCare research patient data repository. The 2018 National NLP Clinical Challenges (n2c2) shared-task workshop presented a cohort selection task for clinical trials [203].

Previously presented NLP tasks and methods were applied to medical data in different languages, with the majority being in English (153/194, 78.9%; Table 2).

Multimedia Appendix 3 presents the CDWs used in the publications presented in this review. Overall, the oldest CDWs, such as the Columbia University Irving Medical Center CDW, Mayo Clinic, and the Partners HealthCare research patient data repository, are the ones that reuse the most textual data and contribute the most to developing the application of NLP on EHR data.

Table 2. Language of the data used in the papers (n=194).

Data language	Publications, n (%)	References
English	153 (78.9)	[15-17,19-25,27-38,41-48,50,51,53-68,71,72,74,75,78-80,83-88,90-92,96,97,99-112,114,116,119-124,126-135,137,140,142-149,151-154,156-159,161,162,165-176,179,181,184-187,189-192,194-196,198,200-208]
French	27 (13.9)	[39,49,52,73,76,77,81,89,93,94,113,115,118,125,136,138,139,155,163,164,177,178,182,183,188,193,197]
German	9 (4.6)	[18,26,69,95,117,150,160,180,199]
Korean	3 (1.5)	[40,65,82]
Japanese	1 (0.5)	[98]
Not mentioned	2 (1)	[70,141]

Discussion

Principal Findings

As CDWs become more prevalent and are adopted in many countries, they open up opportunities for clinical NLP to flourish. This review shows that the use of NLP on data from CDWs is primarily focused on extracting information from clinical texts and identifying patients. Depending on the task, various methods can be used, from symbolic methods to machine learning and deep learning techniques. The oldest CDWs are associated with the most numerous publications. This shows that the use of NLP is not a 1-time event but is intended to be established in the long term. It contributes to the continuous quality improvement of data made available in CDWs.

Symbolic and linguistics methods have still been widely used in recent years, despite the preponderance of deep learning approaches that have shown excellent results across a majority of tasks. This shows that some tasks can be partially achieved with classical NLP techniques, such as regular expressions and pattern matching that exploit specialized lexica such as drug lists and terminologies. Existing information extraction tools such as cTAKES, MedLEE, and MetaMap offer easy handling and satisfactory results. As a result, they are often used for processing English language clinical text.

Interestingly, the number of data languages presented in our review is quite low—only 5 languages: English, French, German, Korean, and Japanese. This can be explained by three factors: (1) CDWs are not cited as data sources in articles, resulting in a bias related to queries; (2) CDWs are operational in another country, but NLP has not yet been used on these data; and (3) CDWs have not yet been adopted in every country.

Opportunities and Challenges

Although NLP methods are becoming increasingly popular, there remain challenges within the clinical field. This review demonstrates that the use of NLP in CDWs is becoming more frequent over time. However, CDWs still rarely provide open access for NLP research owing to medical data confidentiality. A first step to partially overcome the privacy constraints could involve working on deidentified or anonymized data from CDWs, as has been done in some recent shared tasks [202,204,205,210]. These shared tasks, crucial for making advances in medical NLP research, are too scarce, particularly for languages other than English [9]. Providing an appropriate measure to respect patient privacy should encourage collaboration among hospital and NLP research teams and facilitate access to clinical data.

The global movement is toward the structuring and interoperability of clinical data; yet, the finer points of medical reasoning are always expressed in textual reports, and such information cannot always be structured. The increase in NLP approaches applied to clinical data could lead to major advances in clinical research, both to identify the populations of interest and to retrieve relevant information of these patients for clinical research. NLP could also have a positive impact on the daily life of caregivers by speeding up access to information contained in patient EHRs using automated tools for the summarization of patient history. Indeed, caregivers invest a significant amount of time recording information gathered during care delivery in textual reports. Surprisingly, they also dedicate an equivalent amount of time sifting through numerous documents to retrieve this information when needed.

Structured or semistructured data stored in CDWs provide information about patient follow-up and can serve as a valuable resource for developing or enhancing NLP systems. Indeed, temporal data can offer guidance on where the information is most relevant in the text. In addition, other data such as PHI, including names, surnames, and addresses, can be used as a starting point in NLP systems.

Clinical data are a use case for NLP research. They possess the advantage of being accessible in multiple languages owing to the global nature of medical care. This accessibility enhances research efforts focused on multilingualism. Such data are available in abundance, facilitating the acquisition of effective clinical text representations that can be applied in deep neural networks to learn relevant concept models. Clinical data fall within the category of specialized domains or languages

designed for specific purposes. They share certain properties, such as specific knowledge, uses, and discourse. This also entails undertaking specific tasks such as deidentification or anonymization.

The analysis of the literature conducted here highlights the need for further development of CDWs, with a stronger integration of NLP applications throughout the entire data value chain.

Limitations

The NLP tasks identified in this review cover only a small part of all existing NLP tasks in the general domain. These tasks globally reflect the primary needs in clinical research, such as identifying the study population and extracting clinical information for a defined population. Other tasks, such as context analysis and language modeling, have been widely studied in the general domain NLP but are less prevalent in the clinical domain. In recent years, transformer-based approaches have emerged as the state-of-the-art methods for most NLP tasks. However, this review indicates that these methods have not fully spread to the clinical domain. This demonstrates a gap between methods that are well established in the general domain NLP and their adoption in specific domains such as the clinical domain.

This review focuses on 2 very specific subjects from different emerging domains: clinical NLP and CDWs. This combination of subjects implies the use of multiple bibliographic databases and the aggregation of multiple queries to ensure good coverage of the literature. Some bibliographic databases cover a wider range of articles and include articles already present in other more specialized sources. To avoid having a surfeit of duplicate articles, we prioritized the use of the most encompassing bibliographic databases: Google Scholar and PubMed. This introduces a bias of completeness because relevant articles could be missing from the selected bibliographic databases and be present in others we did not use in this review, such as Scopus, Web of Science, and Embase.

There is another bias of completeness related to the search by keywords in the bibliographic databases. A given concept can be expressed in various ways in natural language, using different keywords. The choice of keywords is crucial to aim at both high specificity and high sensitivity, even if the selected keywords are searched in the whole paper. In this review, we used very broad keywords to have the highest sensitivity but at the expense of specificity (n=194, 14.34% relevant articles among 1353 articles identified from the queries).

Conclusions

CDWs are central to the secondary use of clinical texts for research purposes. Our review highlights the growing interest in computerized health data, particularly in clinical texts, where NLP is used to address various clinical tasks. These tasks include patient identification and information extraction, as well as clinical NLP tasks such as language modeling, context analysis, and EHR deidentification. The broad spectrum of NLP approaches has been effectively leveraged, ranging from symbolic methods to machine learning and deep learning methods. Despite the prevalence of pretrained language models in the broader NLP domain, symbolic and linguistics methods have continued to be used in recent years. In the realm of clinical NLP for CDWs, the trends align with global NLP patterns, where resources and methods are predominantly developed for the English language. The development of NLP in the medical field will require cooperation between health care and NLP experts.

Acknowledgments

This work was supported by the French Agence Nationale de la Recherche (ANR; National Research Agency) AIBy4 project (ANR-20-THIA-0011).

Conflicts of Interest

None declared.

Multimedia Appendix 1
PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[PDF File (Adobe PDF File), 50 KB - medinform_v11i1e42477_app1.pdf]

Multimedia Appendix 2
Search queries used in PubMed, Google Scholar, and ACL Anthology to retrieve publications for inclusion in this systematic review.
[DOCX File , 13 KB - medinform_v11i1e42477_app2.docx]

Multimedia Appendix 3
Clinical data warehouses from which data have been used in a publication.
[DOCX File , 21 KB - medinform_v11i1e42477_app3.docx]

References

1. Meystre SM, Lovis C, Bürkle T, Tognola G, Budrionis A, Lehmann CU. Clinical data reuse or secondary use: current status and potential future progress. Yearb Med Inform 2017 Aug;26(1):38-52. [doi: 10.15265/IY-2017-007] [Medline: 28480475]
2. Adler-Milstein J, Holmgren AJ, Kralovec P, Worzala C, Searcy T, Patel V. Electronic health record adoption in US hospitals: the emergence of a digital "advanced use" divide. J Am Med Inform Assoc 2017 Nov 01;24(6):1142-1148 [FREE Full text] [doi: 10.1093/jamia/ocx080] [Medline: 29016973]
3. Casto AB, Layman E. Principles of Healthcare Reimbursement. Springfield, IL: American Health Information Management Association; 2013:371.
4. Köpcke F, Prokosch HU. Employing computers for the recruitment into clinical trials: a comprehensive systematic review. J Med Internet Res 2014 Jul 01;16(7):e161 [FREE Full text] [doi: 10.2196/jmir.3446] [Medline: 24985568]
5. Shah SM, Khan RA. Secondary use of electronic health record: opportunities and challenges. IEEE Access 2020;8:136947-136965. [doi: 10.1109/access.2020.3011099]
6. Sarwar T, Seifollahi S, Chan J, Zhang X, Aksakalli V, Hudson I, et al. The secondary use of electronic health records for data mining: data characteristics and challenges. ACM Comput Surv 2022 Jan 18;55(2):1-40 [FREE Full text] [doi: 10.1145/3490234]
7. Hamoud A, Hashim A, Awadh W. Clinical data warehouse: a review. Iraqi J Comput Inform 2018 Dec 31;44(2):16-26 [FREE Full text] [doi: 10.25195/ijci.v44i2.53]
8. Holmes JH, Elliott TE, Brown JS, Raebel MA, Davidson A, Nelson AF, et al. Clinical research data warehouse governance for distributed research networks in the USA: a systematic review of the literature. J Am Med Inform Assoc 2014 Jul;21(4):730-736 [FREE Full text] [doi: 10.1136/amiajnl-2013-002370] [Medline: 24682495]
9. Gagalova KK, Leon Elizalde MA, Portales-Casamar E, Görges M. What you need to know before implementing a clinical research data warehouse: comparative review of integrated data repositories in health care institutions. JMIR Form Res 2020 Aug 27;4(8):e17687 [FREE Full text] [doi: 10.2196/17687] [Medline: 32852280]
10. Lin WC, Chen JS, Chiang MF, Hribar MR. Applications of artificial intelligence to electronic health record data in ophthalmology. Transl Vis Sci Technol 2020 Feb 27;9(2):13 [FREE Full text] [doi: 10.1167/tvst.9.2.13] [Medline: 32704419]
11. Juhn Y, Liu H. Artificial intelligence approaches using natural language processing to advance EHR-based clinical research. J Allergy Clin Immunol 2020 Feb;145(2):463-469 [FREE Full text] [doi: 10.1016/j.jaci.2019.12.897] [Medline: 31883846]
12. Kim E, Rubinstein SM, Nead KT, Wojcieszynski AP, Gabriel PE, Warner JL. The evolving use of electronic health records (EHR) for research. Semin Radiat Oncol 2019 Oct;29(4):354-361. [doi: 10.1016/j.semradonc.2019.05.010] [Medline: 31472738]
13. Névéol A, Dalianis H, Velupillai S, Savova G, Zweigenbaum P. Clinical natural language processing in languages other than English: opportunities and challenges. J Biomed Semantics 2018 Mar 30;9(1):12 [FREE Full text] [doi: 10.1186/s13326-018-0179-8] [Medline: 29602312]
14. Publish or perish. Anne-Wil Harzing. URL: https://harzing.com/resources/publish-or-perish [accessed 2023-11-27]
15. Yu Z, Yang X, Dang C, Wu S, Adekkanattu P, Pathak J, et al. A study of social and behavioral determinants of health in lung cancer patients using transformers-based natural language processing models. AMIA Annu Symp Proc 2021 Feb 21;2021:1225-1233 [FREE Full text] [Medline: 35309014]
16. Afshar M, Dligach D, Sharma B, Cai X, Boyda J, Birch S, et al. Development and application of a high throughput natural language processing architecture to convert all clinical documents in a clinical data warehouse into standardized medical vocabularies. J Am Med Inform Assoc 2019 Nov 01;26(11):1364-1369 [FREE Full text] [doi: 10.1093/jamia/ocz068] [Medline: 31145455]
17. Raja AS, Pourjabbar S, Ip IK, Baugh CW, Sodickson AD, O'Leary M, et al. Impact of a health information technology-enabled appropriate use criterion on utilization of emergency department CT for renal colic. AJR Am J Roentgenol 2019 Jan;212(1):142-145. [doi: 10.2214/AJR.18.19966] [Medline: 30403534]
18. Grön L, Bertels A, Heylen K. Leveraging sublanguage features for the semantic categorization of clinical terms. In: Proceedings of the 18th BioNLP Workshop and Shared Task. 2019 Presented at: BioNLP '19; August 1, 2019; Florence, Italy p. 211-216 URL: https://aclanthology.org/W19-5022.pdf [doi: 10.18653/v1/w19-5022]
19. Wang L, Haug PJ, Del Fiol G. Using classification models for the generation of disease-specific medications from biomedical literature and clinical data repository. J Biomed Inform 2017 May;69:259-266 [FREE Full text] [doi: 10.1016/j.jbi.2017.04.014] [Medline: 28435015]
20. Walsh JA, Shao Y, Leng J, He T, Teng CC, Redd D, et al. Identifying axial spondyloarthritis in electronic medical records of US veterans. Arthritis Care Res (Hoboken) 2017 Sep;69(9):1414-1420. [doi: 10.1002/acr.23140] [Medline: 27813310]
21. Perotte A, Wood F, Elhadad N, Wood F. Hierarchically supervised Latent Dirichlet allocation. In: Proceedings of the 24th International Conference on Neural Information Processing Systems. 2011 Presented at: NIPS '11; December 12-15, 2011; Granada, Spain p. 2609-2617 URL: https://dl.acm.org/doi/10.5555/2986459.2986750
22. Zhong QY, Karlson EW, Gelaye B, Finan S, Avillach P, Smoller JW, et al. Screening pregnant women for suicidal behavior in electronic medical records: diagnostic codes vs. clinical notes processed by natural language processing. BMC Med Inform Decis Mak 2018 May 29;18(1):30 [FREE Full text] [doi: 10.1186/s12911-018-0617-7] [Medline: 29843698]
23. Jonnalagadda S, Cohen T, Wu S, Gonzalez G. Enhancing clinical concept extraction with distributional semantics. J Biomed Inform 2012 Feb;45(1):129-140 [FREE Full text] [doi: 10.1016/j.jbi.2011.10.007] [Medline: 22085698]

24. Chase HS, Mitrani LR, Lu GG, Fulgieri DJ. Early recognition of multiple sclerosis using natural language processing of the electronic health record. BMC Med Inform Decis Mak 2017 Feb 28;17(1):24 [FREE Full text] [doi: 10.1186/s12911-017-0418-4] [Medline: 28241760]
25. Ashish N, Dahm L, Boicey C. University of California, Irvine-Pathology Extraction Pipeline: the pathology extraction pipeline for information extraction from pathology reports. Health Informatics J 2014 Dec;20(4):288-305 [FREE Full text] [doi: 10.1177/1460458213494032] [Medline: 25155030]
26. Scheurwegs E, Luyckx K, Luyten L, Goethals B, Daelemans W. Assigning clinical codes with data-driven concept representation on Dutch clinical free text. J Biomed Inform 2017 May;69:118-127 [FREE Full text] [doi: 10.1016/j.jbi.2017.04.007] [Medline: 28400312]
27. Melton GB, Parsons S, Morrison FP, Rothschild AS, Markatou M, Hripcsak G. Inter-patient distance metrics using SNOMED CT defining relationships. J Biomed Inform 2006 Dec;39(6):697-705 [FREE Full text] [doi: 10.1016/j.jbi.2006.01.004] [Medline: 16554186]
28. Wang X, Chused A, Elhadad N, Friedman C, Markatou M. Automated knowledge acquisition from clinical narrative reports. AMIA Annu Symp Proc 2008 Nov 06;2008:783-787 [FREE Full text] [Medline: 18999156]
29. Wang X, Hripcsak G, Markatou M, Friedman C. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. J Am Med Inform Assoc 2009;16(3):328-337 [FREE Full text] [doi: 10.1197/jamia.M3028] [Medline: 19261932]
30. Lowe HJ, Huang Y, Regula DP. Using a statistical natural language parser augmented with the UMLS specialist lexicon to assign SNOMED CT codes to anatomic sites and pathologic diagnoses in full text pathology reports. AMIA Annu Symp Proc 2009 Nov 14;2009:386-390 [FREE Full text] [Medline: 20351885]
31. Harris DR, Henderson DW, Corbeau A. sig2db: a workflow for processing natural language from prescription instructions for clinical data warehouses. AMIA Jt Summits Transl Sci Proc 2020 May 30;2020:221-230 [FREE Full text] [Medline: 32477641]
32. Chuang JH, Friedman C, Hripcsak G. A comparison of the Charlson comorbidities derived from medical language processing and administrative data. Proc AMIA Symp 2002:160-164 [FREE Full text] [Medline: 12463807]
33. Van Vleck TT, Wilcox A, Stetson PD, Johnson SB, Elhadad N. Content and structure of clinical problem lists: a corpus analysis. AMIA Annu Symp Proc 2008 Nov 06;2008:753-757 [FREE Full text] [Medline: 18999284]
34. Li L, Chase HS, Patel CO, Friedman C, Weng C. Comparing ICD9-encoded diagnoses and NLP-processed discharge summaries for clinical trials pre-screening: a case study. AMIA Annu Symp Proc 2008 Nov 06;2008:404-408 [FREE Full text] [Medline: 18999285]
35. Chen ES, Hripcsak G, Xu H, Markatou M, Friedman C. Automated acquisition of disease drug knowledge from biomedical and clinical documents: an initial study. J Am Med Inform Assoc 2008;15(1):87-98 [FREE Full text] [doi: 10.1197/jamia.M2401] [Medline: 17947625]
36. Carlo L, Chase HS, Weng C. Aligning structured and unstructured medical problems using UMLS. AMIA Annu Symp Proc 2010 Nov 13;2010:91-95 [FREE Full text] [Medline: 21346947]
37. Zhou X, Wang Y, Sohn S, Therneau TM, Liu H, Knopman DS. Automatic extraction and assessment of lifestyle exposures for Alzheimer's disease using natural language processing. Int J Med Inform 2019 Oct;130:103943 [FREE Full text] [doi: 10.1016/j.ijmedinf.2019.08.003] [Medline: 31476655]
38. Singh K, Betensky RA, Wright A, Curhan GC, Bates DW, Waikar SS. A concept-wide association study of clinical notes to discover new predictors of kidney failure. Clin J Am Soc Nephrol 2016 Dec 07;11(12):2150-2158 [FREE Full text] [doi: 10.2215/CJN.02420316] [Medline: 27927892]
39. Campillo-Gimenez B, Garcelon N, Jarno P, Chapplain JM, Cuggia M. Full-text automated detection of surgical site infections secondary to neurosurgery in Rennes, France. Stud Health Technol Inform 2013;192:572-575. [Medline: 23920620]
40. Hong SN, Son HJ, Choi SK, Chang DK, Kim Y, Jung S, et al. A prediction model for advanced colorectal neoplasia in an asymptomatic screening population. PLoS One 2017;12(8):e0181040 [FREE Full text] [doi: 10.1371/journal.pone.0181040] [Medline: 28841657]
41. Hunter-Zinck HS, Peck JS, Strout TD, Gaehde SA. Predicting emergency department orders with multilabel machine learning techniques and simulating effects on length of stay. J Am Med Inform Assoc 2019 Dec 01;26(12):1427-1436 [FREE Full text] [doi: 10.1093/jamia/ocz171] [Medline: 31578568]
42. Kshatriya BS, Balls-Berry JE, Freeman WD, Zhang R, Wang Y. Completeness of Social and Behavioral Determinants of Health in Electronic Health Records: A case study on the Patient-Provided Information from a minority cohort with sexually transmitted diseases. Research Square. Preprint posted online December 10, 2020 2020 [FREE Full text] [doi: 10.21203/rs.3.rs-123744/v1]
43. Baghal A, Al-Shukri S, Kumari A. Agile natural language processing model for pathology knowledge extraction and integration with clinical enterprise data warehouse. In: Proceedings of the 6th International Conference on Social Networks Analysis, Management and Security. 2019 Presented at: SNAMS '19; October 22-25, 2019; Granada, Spain p. 419-422 URL: https://ieeexplore.ieee.org/document/8931828 [doi: 10.1109/snams.2019.8931828]
44. Liu H, Wu ST, Li D, Jonnalagadda S, Sohn S, Wagholikar K, et al. Towards a semantic lexicon for clinical natural language processing. AMIA Annu Symp Proc 2012;2012:568-576 [FREE Full text] [Medline: 23304329]

45. Afzal N, Sohn S, Abram S, Scott CG, Chaudhry R, Liu H, et al. Mining peripheral arterial disease cases from narrative clinical notes using natural language processing. J Vasc Surg 2017 Jun;65(6):1753-1761 [FREE Full text] [doi: 10.1016/j.jvs.2016.11.031] [Medline: 28189359]
46. Jonnalagadda S, Cohen T, Wu S, Liu H, Gonzalez G. Using empirically constructed lexical resources for named entity recognition. Biomed Inform Insights 2013 Jun 24;6(Suppl 1):17-27 [FREE Full text] [doi: 10.4137/BII.S11664] [Medline: 23847424]
47. Osborne JD, Wyatt M, Westfall AO, Willig J, Bethard S, Gordon G. Efficient identification of nationally mandated reportable cancer cases using natural language processing and machine learning. J Am Med Inform Assoc 2016 Nov;23(6):1077-1084 [FREE Full text] [doi: 10.1093/jamia/ocw006] [Medline: 27026618]
48. Hernandez-Boussard T, Blayney DW, Brooks JD. Leveraging digital data to inform and improve quality cancer care. Cancer Epidemiol Biomarkers Prev 2020 Apr;29(4):816-822 [FREE Full text] [doi: 10.1158/1055-9965.EPI-19-0873] [Medline: 32066619]
49. Lerner I, Paris N, Tannier X. Terminologies augmented recurrent neural network model for clinical named entity recognition. J Biomed Inform 2020 Feb;102:103356 [FREE Full text] [doi: 10.1016/j.jbi.2019.103356] [Medline: 31837473]
50. Wang X, Chase H, Markatou M, Hripcsak G, Friedman C. Selecting information in electronic health records for knowledge acquisition. J Biomed Inform 2010 Aug;43(4):595-601 [FREE Full text] [doi: 10.1016/j.jbi.2010.03.011] [Medline: 20362071]
51. Overby CL, Pathak J, Gottesman O, Haerian K, Perotte A, Murphy S, et al. A collaborative approach to developing an electronic health record phenotyping algorithm for drug-induced liver injury. J Am Med Inform Assoc 2013 Dec;20(e2):e243-e252 [FREE Full text] [doi: 10.1136/amiajnl-2013-001930] [Medline: 23837993]
52. Neuraz A, Lerner I, Digan W, Paris N, Tsopra R, Rogier A, AP-HP/Universities/INSERM COVID-19 Research Collaboration; AP-HP COVID CDR Initiative. Natural language processing for rapid response to emergent diseases: case study of calcium channel blockers and hypertension in the COVID-19 pandemic. J Med Internet Res 2020 Aug 14;22(8):e20773 [FREE Full text] [doi: 10.2196/20773] [Medline: 32759101]
53. Eyuboglu S, Angus G, Patel BN, Pareek A, Davidzon G, Long J, et al. Multi-task weak supervision enables anatomically-resolved abnormality detection in whole-body FDG-PET/CT. Nat Commun 2021 Mar 25;12(1):1880 [FREE Full text] [doi: 10.1038/s41467-021-22018-1] [Medline: 33767174]
54. He T, Puppala M, Ezeana CF, Huang Y, Chou P, Yu X, et al. A deep learning-based decision support tool for precision risk assessment of breast cancer. JCO Clin Cancer Inform 2019 May;3:1-12 [FREE Full text] [doi: 10.1200/CCI.18.00121] [Medline: 31141423]
55. Golas SB, Shibahara T, Agboola S, Otaki H, Sato J, Nakae T, et al. A machine learning model to predict the risk of 30-day readmissions in patients with heart failure: a retrospective analysis of electronic medical records data. BMC Med Inform Decis Mak 2018 Jun 22;18(1):44 [FREE Full text] [doi: 10.1186/s12911-018-0620-z] [Medline: 29929496]
56. Sehdev A, Hayden R, Kuhar MJ, Cheng L, Warren SJ, Mark LA, et al. Prognostic role of BRAF mutation in malignant cutaneous melanoma. J Clin Oncol 2018 May 20;36(15_suppl):e21599 [FREE Full text] [doi: 10.1200/jco.2018.36.15_suppl.e21599]
57. Riestenberg RA, Furman A, Cowen A, Pawlowksi A, Schneider D, Lewis AA, et al. Differences in statin utilization and lipid lowering by race, ethnicity, and HIV status in a real-world cohort of persons with human immunodeficiency virus and uninfected persons. Am Heart J 2019 Mar;209:79-87 [FREE Full text] [doi: 10.1016/j.ahj.2018.11.012] [Medline: 30685678]
58. Abboud A, Nguonly A, Bean A, Brown KJ, Chen RF, Dudzinski D, et al. Rationale and design of the preserved versus reduced ejection fraction biomarker registry and precision medicine database for ambulatory patients with heart failure (PREFER-HF) study. Open Heart 2021 Oct;8(2):e001704 [FREE Full text] [doi: 10.1136/openhrt-2021-001704] [Medline: 34663746]
59. Wang L, Xue Z, Ezeana CF, Puppala M, Chen S, Danforth RL, et al. Preventing inpatient falls with injuries using integrative machine learning prediction: a cohort study. NPJ Digit Med 2019;2:127 [FREE Full text] [doi: 10.1038/s41746-019-0200-3] [Medline: 31872067]
60. Shoenbill K, Song Y, Craven M, Johnson H, Smith M, Mendonca EA. Identifying patterns and predictors of lifestyle modification in electronic health record documentation using statistical and machine learning methods. Prev Med 2020 Jul;136:106061 [FREE Full text] [doi: 10.1016/j.ypmed.2020.106061] [Medline: 32179026]
61. Feller DJ, Zucker J, Yin MT, Gordon P, Elhadad N. Using clinical notes and natural language processing for automated HIV risk assessment. J Acquir Immune Defic Syndr 2018 Feb 01;77(2):160-166 [FREE Full text] [doi: 10.1097/QAI.0000000000001580] [Medline: 29084046]
62. Bozkurt S, Kan KM, Ferrari MK, Rubin DL, Blayney DW, Hernandez-Boussard T, et al. Is it possible to automatically assess pretreatment digital rectal examination documentation using natural language processing? A single-centre retrospective study. BMJ Open 2019 Jul 18;9(7):e027182 [FREE Full text] [doi: 10.1136/bmjopen-2018-027182] [Medline: 31324681]
63. Perotte A, Ranganath R, Hirsch JS, Blei D, Elhadad N. Risk prediction for chronic kidney disease progression using heterogeneous electronic health record data and time series analysis. J Am Med Inform Assoc 2015 Jul;22(4):872-880 [FREE Full text] [doi: 10.1093/jamia/ocv024] [Medline: 25896647]

64. Klang E, Kummer BR, Dangayach NS, Zhong A, Kia MA, Timsina P, et al. Predicting adult neuroscience intensive care unit admission from emergency department triage using a retrospective, tabular-free text machine learning approach. Sci Rep 2021 Jan 14;11(1):1381 [FREE Full text] [doi: 10.1038/s41598-021-80985-3] [Medline: 33446890]
65. Bae YS, Kim KH, Kim HK, Choi SW, Ko T, Seo HH, et al. Keyword extraction algorithm for classifying smoking status from unstructured bilingual electronic health records based on natural language processing. Appl Sci 2021 Sep 22;11(19):8812 [FREE Full text] [doi: 10.3390/app11198812]
66. Stemerman R, Arguello J, Brice J, Krishnamurthy A, Houston M, Kitzmiller R. Identification of social determinants of health using multi-label classification of electronic health record clinical notes. JAMIA Open 2021 Jul;4(3):ooaa069 [FREE Full text] [doi: 10.1093/jamiaopen/ooaa069] [Medline: 34514351]
67. Sharperson C, Hanna TN, Herr KD, Zygmont ME, Gerard RL, Johnson J. The effect of COVID-19 on emergency department imaging: what can we learn? Emerg Radiol 2021 Apr;28(2):339-347 [FREE Full text] [doi: 10.1007/s10140-020-01889-9] [Medline: 33420529]
68. Moon S, Wen A, Scott C. An automated system for analysis of implantable cardioverter defibrillator reports in hypertrophic cardiomyopathy patients. Circulation 2018;138(Suppl 1):A16215. [doi: 10.26226/morressier.5d19cfb257558b317a10dd93]
69. Zubke M, Katzensteiner M, Bott OJ. Stud Health Technol Inform 2020 Jun 16;270:272-276. [doi: 10.3233/SHTI200165] [Medline: 32570389]
70. Ciofi Degli Atti ML, Pecoraro F, Piga S, Luzi D, Raponi M. Developing a surgical site infection surveillance system based on hospital unstructured clinical notes and text mining. Surg Infect (Larchmt) 2020 Oct;21(8):716-721. [doi: 10.1089/sur.2019.238] [Medline: 32105569]
71. Cochon LR, Giess CS, Khorasani R. Comparing diagnostic performance of digital breast tomosynthesis and full-field digital mammography. J Am Coll Radiol 2020 Aug;17(8):999-1003. [doi: 10.1016/j.jacr.2020.01.010] [Medline: 32068009]
72. Lacson R, Wang A, Cochon L, Giess C, Desai S, Eappen S, et al. Factors associated with optimal follow-up in women with BI-RADS 3 breast findings. J Am Coll Radiol 2020 Apr;17(4):469-474 [FREE Full text] [doi: 10.1016/j.jacr.2019.10.003] [Medline: 31669081]
73. Chrusciel J, Girardon F, Roquette L, Laplanche D, Duclos A, Sanchez S. The prediction of hospital length of stay using unstructured data. BMC Med Inform Decis Mak 2021 Dec 18;21(1):351 [FREE Full text] [doi: 10.1186/s12911-021-01722-4] [Medline: 34922532]
74. Stein DM, Vawdrey DK, Stetson PD, Bakken S. An analysis of team checklists in physician signout notes. AMIA Annu Symp Proc 2010 Nov 13;2010:767-771 [FREE Full text] [Medline: 21347082]
75. Chen W, Butler RK, Zhou Y, Parker RA, Jeon CY, Wu BU. Prediction of pancreatic cancer based on imaging features in patients with duct abnormalities. Pancreas 2020 Mar;49(3):413-419 [FREE Full text] [doi: 10.1097/MPA.0000000000001499] [Medline: 32132511]
76. Rogier A, Coulet A, Rance B. Using an ontological representation of chemotherapy toxicities for guiding information extraction and integration from EHRs. Stud Health Technol Inform 2022 Jun 06;290:91-95. [doi: 10.3233/SHTI220038] [Medline: 35672977]
77. Delespierre T, Denormandie P, Bar-Hen A, Josseran L. Empirical advances with text mining of electronic health records. BMC Med Inform Decis Mak 2017 Aug 22;17(1):127 [FREE Full text] [doi: 10.1186/s12911-017-0519-0] [Medline: 28830417]
78. Wang L, Wampfler J, Dispenzieri A, Xu H, Yang P, Liu H. Achievability to extract specific date information for cancer research. AMIA Annu Symp Proc 2019;2019:893-902 [FREE Full text] [Medline: 32308886]
79. Genes N, Chandra D, Ellis S, Baumlin K. Validating emergency department vital signs using a data quality engine for data warehouse. Open Med Inform J 2013;7:34-39 [FREE Full text] [doi: 10.2174/1874431101307010034] [Medline: 24403981]
80. Glaser AP, Jordan BJ, Cohen J, Desai A, Silberman P, Meeks JJ. Automated extraction of grade, stage, and quality information from transurethral resection of bladder tumor pathology reports using natural language processing. JCO Clin Cancer Inform 2018 Dec;2:1-8 [FREE Full text] [doi: 10.1200/CCI.17.00128] [Medline: 30652586]
81. Duthe JC, Bouzille G, Sylvestre E, Chazard E, Arvieux C, Cuggia M. How to identify potential candidates for HIV Pre-exposure prophylaxis: an AI algorithm reusing real-world hospital data. Stud Health Technol Inform 2021 May 27;281:714-718. [doi: 10.3233/SHTI210265] [Medline: 34042669]
82. Lee KH, Kim HJ, Kim YJ, Kim JH, Song EY. Extracting structured genotype information from free-text HLA reports using a rule-based approach. J Korean Med Sci 2020 Mar 30;35(12):e78 [FREE Full text] [doi: 10.3346/jkms.2020.35.e78] [Medline: 32233158]
83. Tamang S, Patel MI, Blayney DW, Kuznetsov J, Finlayson SG, Vetteth Y, et al. Detecting unplanned care from clinician notes in electronic health records. J Oncol Pract 2015 May;11(3):e313-e319 [FREE Full text] [doi: 10.1200/JOP.2014.002741] [Medline: 25980019]
84. Yang X, Yang H, Lyu T, Yang S, Guo Y, Bian J, et al. A natural language processing tool to extract quantitative smoking status from clinical narratives. IEEE Int Conf Healthc Inform 2020;2020:1109 [FREE Full text] [doi: 10.1109/ICHI48887.2020.9374369] [Medline: 33786419]

85. Wang Y, Mehrabi S, Sohn S, Atkinson EJ, Amin S, Liu H. Natural language processing of radiology reports for identification of skeletal site-specific fractures. BMC Med Inform Decis Mak 2019 Apr 04;19(Suppl 3):73 [FREE Full text] [doi: 10.1186/s12911-019-0780-5] [Medline: 30943952]
86. Zhao Y, Weroha SJ, Goode EL, Liu H, Wang C. Generating real-world evidence from unstructured clinical notes to examine clinical utility of genetic tests: use case in BRCAness. BMC Med Inform Decis Mak 2021 Jan 06;21(1):3 [FREE Full text] [doi: 10.1186/s12911-020-01364-y] [Medline: 33407429]
87. Haug PJ, Ferraro JP, Holmen J, Wu X, Mynam K, Ebert M, et al. An ontology-driven, diagnostic modeling system. J Am Med Inform Assoc 2013 Jun;20(e1):e102-e110 [FREE Full text] [doi: 10.1136/amiajnl-2012-001376] [Medline: 23523876]
88. Magnani CJ, Bievre N, Baker LC, Brooks JD, Blayney DW, Hernandez-Boussard T. Real-world evidence to estimate prostate cancer costs for first-line treatment or active surveillance. Eur Urol Open Sci 2021 Jan;23:20-29 [FREE Full text] [doi: 10.1016/j.euros.2020.11.004] [Medline: 33367287]
89. Ansoborlo M, Dhalluin T, Gaborit C, Cuggia M, Grammatico-Guilllon L. Prescreening in oncology using data sciences: the PreScIOUS study. Stud Health Technol Inform 2021 May 27;281:123-127. [doi: 10.3233/SHTI210133] [Medline: 34042718]
90. Ryu JH, Zimolzak AJ. Natural language processing of serum protein electrophoresis reports in the veterans affairs health care system. JCO Clin Cancer Inform 2020 Aug;4:749-756 [FREE Full text] [doi: 10.1200/CCI.19.00167] [Medline: 32813561]
91. Luther SL, Thomason SS, Sabharwal S, Finch DK, McCart J, Toyinbo P, et al. Leveraging electronic health care record information to measure pressure ulcer risk in veterans with spinal cord injury: a longitudinal study protocol. JMIR Res Protoc 2017 Jan 19;6(1):e3 [FREE Full text] [doi: 10.2196/resprot.5948] [Medline: 28104580]
92. Liu Y, Lependu P, Iyer S, Shah NH. Using temporal patterns in medical records to discern adverse drug events from indications. AMIA Jt Summits Transl Sci Proc 2012;2012:47-56 [FREE Full text] [Medline: 22779050]
93. Lerner I, Jouffroy J, Burgun A. Learning the grammar of drug prescription: recurrent neural network grammars for medication information extraction in clinical texts. arXiv. Preprint posted online April 24, 2020 2020 [FREE Full text] [doi: 10.48550/arXiv.2004.11622]
94. Hoertel N, Sánchez-Rico M, Vernet R, Beeker N, Neuraz A, Alvarado JM, AP-HP/Université de Paris/INSERM Covid-19 research collaborationAP-HP Covid CDR Initiative. Dexamethasone use and mortality in hospitalized patients with coronavirus disease 2019: a multicentre retrospective observational study. Br J Clin Pharmacol 2021 Oct;87(10):3766-3775 [FREE Full text] [doi: 10.1111/bcp.14784] [Medline: 33608891]
95. Caliskan D, Zierk J, Kraska D, Schulz S, Daumke P, Prokosch HU, et al. First steps to evaluate an NLP tool's medication extraction accuracy from discharge letters. Stud Health Technol Inform 2021 May 24;278:224-230. [doi: 10.3233/SHTI210073] [Medline: 34042898]
96. Rochefort CM, Buckeridge DL, Abrahamowicz M. Improving patient safety by optimizing the use of nursing human resources. Implement Sci 2015 Jun 14;10(1):89 [FREE Full text] [doi: 10.1186/s13012-015-0278-1] [Medline: 26071752]
97. Wang G, Jung K, Winnenburg R, Shah NH. A method for systematic discovery of adverse drug events from clinical notes. J Am Med Inform Assoc 2015 Nov;22(6):1196-1204 [FREE Full text] [doi: 10.1093/jamia/ocv102] [Medline: 26232442]
98. Shimai Y, Takeda T, Okada K, Manabe S, Teramoto K, Mihara N, et al. Screening of anticancer drugs to detect drug-induced interstitial pneumonia using the accumulated data in the electronic medical record. Pharmacol Res Perspect 2018 Jul;6(4):e00421 [FREE Full text] [doi: 10.1002/prp2.421] [Medline: 30009034]
99. Jung K, Lependu P, Shah N. Automated detection of systematic off-label drug use in free text of electronic medical records. AMIA Jt Summits Transl Sci Proc 2013;2013:94-98 [FREE Full text] [Medline: 24303308]
100. Geva A, Abman S, Manzi S, Ivy DD, Mullen MP, Griffin J, et al. Adverse drug event rates in pediatric pulmonary hypertension: a comparison of real-world data sources. J Am Med Inform Assoc 2020 Feb 01;27(2):294-300 [FREE Full text] [doi: 10.1093/jamia/ocz194] [Medline: 31769835]
101. Rochefort CM, Buckeridge DL, Tanguay A, Biron A, D'Aragon F, Wang S, et al. Accuracy and generalizability of using automated methods for identifying adverse events from electronic health record data: a validation study protocol. BMC Health Serv Res 2017 Feb 16;17(1):147 [FREE Full text] [doi: 10.1186/s12913-017-2069-7] [Medline: 28209197]
102. Gold S, Elhadad N, Zhu X, Cimino JJ, Hripcsak G. Extracting structured medication event information from discharge summaries. AMIA Annu Symp Proc 2008 Nov 06;2008:237-241 [FREE Full text] [Medline: 18999147]
103. Li Y, Salmasian H, Harpaz R, Chase H, Friedman C. Determining the reasons for medication prescriptions in the EHR using knowledge and natural language processing. AMIA Annu Symp Proc 2011;2011:768-776 [FREE Full text] [Medline: 22195134]
104. Harpaz R, Vilar S, Dumouchel W, Salmasian H, Haerian K, Shah NH, et al. Combing signals from spontaneous reports and electronic health records for detection of adverse drug reactions. J Am Med Inform Assoc 2013 May 01;20(3):413-419 [FREE Full text] [doi: 10.1136/amiajnl-2012-000930] [Medline: 23118093]
105. Zhao Y, Dimou A, Shen F, Zong N, Davila JI, Liu H, et al. PO2RDF: representation of real-world data for precision oncology using resource description framework. BMC Med Genomics 2022 Jul 30;15(1):167 [FREE Full text] [doi: 10.1186/s12920-022-01314-9] [Medline: 35907849]

106. Lependu P, Liu Y, Iyer S, Udell MR, Shah NH. Analyzing patterns of drug use in clinical notes for patient safety. AMIA Jt Summits Transl Sci Proc 2012;2012:63-70 [FREE Full text] [Medline: 22779054]
107. Lependu P, Iyer SV, Fairon C, Shah NH. Annotation analysis for testing drug safety signals using unstructured clinical notes. J Biomed Semantics 2012 Apr 24;3 Suppl 1(Suppl 1):S5 [FREE Full text] [doi: 10.1186/2041-1480-3-S1-S5] [Medline: 22541596]
108. LePendu P, Iyer SV, Bauer-Mehren A, Harpaz R, Mortensen JM, Podchiyska T, et al. Pharmacovigilance using clinical notes. Clin Pharmacol Ther 2013 Jun;93(6):547-555 [FREE Full text] [doi: 10.1038/clpt.2013.47] [Medline: 23571773]
109. Jung K, LePendu P, Iyer S, Bauer-Mehren A, Percha B, Shah NH. Functional evaluation of out-of-the-box text-mining tools for data-mining tasks. J Am Med Inform Assoc 2015 Jan;22(1):121-131 [FREE Full text] [doi: 10.1136/amiajnl-2014-002902] [Medline: 25336595]
110. Wright A, McCoy A, Henkin S, Flaherty M, Sittig D. Validation of an association rule mining-based method to infer associations between medications and problems. Appl Clin Inform 2013;4(1):100-109 [FREE Full text] [doi: 10.4338/ACI-2012-12-RA-0051] [Medline: 23650491]
111. Malec SA, Wei P, Bernstam EV, Boyce RD, Cohen T. Using computable knowledge mined from the literature to elucidate confounders for EHR-based pharmacovigilance. J Biomed Inform 2021 May;117:103719 [FREE Full text] [doi: 10.1016/j.jbi.2021.103719] [Medline: 33716168]
112. Weeks HL, Beck C, McNeer E, Williams ML, Bejan CA, Denny JC, et al. medExtractR: a targeted, customizable approach to medication extraction from electronic health records. J Am Med Inform Assoc 2020 Mar 01;27(3):407-418 [FREE Full text] [doi: 10.1093/jamia/ocz207] [Medline: 31943012]
113. Chouchana L, Beeker N, Garcelon N, Rance B, Paris N, Salamanca E, AP-HP/Universities/Inserm COVID-19 research collaboration, AP-HP Covid CDR Initiative,"Entrepôt de Données de Santé" AP-HP Consortium". Association of antihypertensive agents with the risk of in-hospital death in patients with COVID-19. Cardiovasc Drugs Ther 2022 Jun;36(3):483-488 [FREE Full text] [doi: 10.1007/s10557-021-07155-5] [Medline: 33595761]
114. Leeper NJ, Bauer-Mehren A, Iyer SV, Lependu P, Olson C, Shah NH. Practice-based evidence: profiling the safety of cilostazol by text-mining of clinical notes. PLoS One 2013;8(5):e63499 [FREE Full text] [doi: 10.1371/journal.pone.0063499] [Medline: 23717437]
115. Jouffroy J, Feldman SF, Lerner I, Rance B, Burgun A, Neuraz A. Hybrid deep learning for medication-related information extraction from clinical texts in french: MedExt algorithm development study. JMIR Med Inform 2021 Mar 16;9(3):e17934 [FREE Full text] [doi: 10.2196/17934] [Medline: 33724196]
116. Min TL, Xu L, Choi JD, Hu R, Allen JW, Reeves C, et al. COVID-19 pandemic-associated changes in the acuity of brain MRI findings: a secondary analysis of reports using natural language processing. Curr Probl Diagn Radiol 2022;51(4):529-533 [FREE Full text] [doi: 10.1067/j.cpradiol.2021.11.001] [Medline: 34955284]
117. Fiebeck J, Laser H, Winther HB, Gerbel S. Leaving no stone unturned: using machine learning based approaches for information extraction from full texts of a research data warehouse. In: Proceedings of the 13th International Conference on Data Integration in the Life Sciences. 2018 Presented at: DILS '18; November 20-21, 2018; Hannover, Germany p. 50-58 URL: https://link.springer.com/chapter/10.1007/978-3-030-06016-9_5 [doi: 10.1007/978-3-030-06016-9_5]
118. Pham AD, Névéol A, Lavergne T, Yasunaga D, Clément O, Meyer G, et al. Natural language processing of radiology reports for the detection of thromboembolic diseases and clinically relevant incidental findings. BMC Bioinformatics 2014 Aug 07;15(1):266 [FREE Full text] [doi: 10.1186/1471-2105-15-266] [Medline: 25099227]
119. Chokshi FH, Shin B, Lee T. Natural language processing for classification of acute, communicable findings on unstructured head CT reports: comparison of neural network and non-neural machine learning techniques. bioRxiv. Preprint posted online August 10, 2017 2017 [FREE Full text] [doi: 10.1101/173310]
120. Cao H, Markatou M, Melton GB, Chiang MF, Hripcsak G. Mining a clinical data warehouse to discover disease-finding associations using co-occurrence statistics. AMIA Annu Symp Proc 2005;2005:106-110 [FREE Full text] [Medline: 16779011]
121. Patel TA, Puppala M, Ogunti RO, Ensor JE, He T, Shewale JB, et al. Correlating mammographic and pathologic findings in clinical decision support using natural language processing and data mining methods. Cancer 2017 Jan 01;123(1):114-121 [FREE Full text] [doi: 10.1002/cncr.30245] [Medline: 27571243]
122. Olmsted ZT, Hadanny A, Marchese AM, DiMarzio M, Khazen O, Argoff C, et al. Recommendations for neuromodulation in diabetic neuropathic pain. Front Pain Res (Lausanne) 2021 Sep 07;2:726308 [FREE Full text] [doi: 10.3389/fpain.2021.726308] [Medline: 35295414]
123. He T, Puppala M, Ogunti R. Deep learning analytics for diagnostic support of breast cancer disease management. In: Proceedings of the 2017 IEEE EMBS International Conference on Biomedical & Health Informatic. 2017 Presented at: BHI '17; February 16-19, 2017; Orlando, FL p. 365-368 URL: https://ieeexplore.ieee.org/document/7897281 [doi: 10.1109/bhi.2017.7897281]
124. He T, Fong JN, Moore LW, Ezeana CF, Victor D, Divatia M, et al. An imageomics and multi-network based deep learning model for risk assessment of liver transplantation for hepatocellular cancer. Comput Med Imaging Graph 2021 Apr;89:101894 [FREE Full text] [doi: 10.1016/j.compmedimag.2021.101894] [Medline: 33725579]

125. Lo Barco T, Kuchenbuch M, Garcelon N, Neuraz A, Nabbout R. Improving early diagnosis of rare diseases using natural language processing in unstructured medical records: an illustration from Dravet syndrome. Orphanet J Rare Dis 2021 Jul 13;16(1):309 [FREE Full text] [doi: 10.1186/s13023-021-01936-9] [Medline: 34256808]
126. Alba PR, Gao A, Lee KM, Anglin-Foote T, Robison B, Katsoulakis E, et al. Ascertainment of veterans with metastatic prostate cancer in electronic health records: demonstrating the case for natural language processing. JCO Clin Cancer Inform 2021 Sep;5:1005-1014 [FREE Full text] [doi: 10.1200/CCI.21.00030] [Medline: 34570630]
127. Zhu VJ, Lenert LA, Bunnell BE, Obeid JS, Jefferson M, Halbert CH. Automatically identifying social isolation from clinical narratives for patients with prostate Cancer. BMC Med Inform Decis Mak 2019 Mar 14;19(1):43 [FREE Full text] [doi: 10.1186/s12911-019-0795-y] [Medline: 30871518]
128. To D, Sharma B, Karnik N, Joyce C, Dligach D, Afshar M. Validation of an alcohol misuse classifier in hospitalized patients. Alcohol 2020 May;84:49-55. [doi: 10.1016/j.alcohol.2019.09.008] [Medline: 31574300]
129. Zong N, Ngo V, Stone DJ, Wen A, Zhao Y, Yu Y, et al. Leveraging genetic reports and electronic health records for the prediction of primary cancers: algorithm development and validation study. JMIR Med Inform 2021 May 25;9(5):e23586 [FREE Full text] [doi: 10.2196/23586] [Medline: 34032581]
130. De Freitas JK, Johnson KW, Golden E, Nadkarni GN, Dudley JT, Bottinger EP, et al. Phe2vec: automated disease phenotyping based on unsupervised embeddings from electronic health records. Patterns (N Y) 2021 Sep 10;2(9):100337 [FREE Full text] [doi: 10.1016/j.patter.2021.100337] [Medline: 34553174]
131. Carter GC, Landsman-Blumberg PB, Johnson BH, Juneau P, Nicol SJ, Li L, et al. KRAS testing of patients with metastatic colorectal cancer in a community-based oncology setting: a retrospective database analysis. J Exp Clin Cancer Res 2015 Mar 27;34(1):29 [FREE Full text] [doi: 10.1186/s13046-015-0146-5] [Medline: 25888436]
132. Banda JM, Halpern Y, Sontag D, Shah NH. Electronic phenotyping with APHRODITE and the Observational Health Sciences and Informatics (OHDSI) data network. AMIA Jt Summits Transl Sci Proc 2017;2017:48-57 [FREE Full text] [Medline: 28815104]
133. Shao Y, Zeng QT, Chen KK, Shutes-David A, Thielke SM, Tsuang DW. Detection of probable dementia cases in undiagnosed patients using structured and unstructured electronic health records. BMC Med Inform Decis Mak 2019 Jul 09;19(1):128 [FREE Full text] [doi: 10.1186/s12911-019-0846-4] [Medline: 31288818]
134. Sharma B, Dligach D, Swope K, Salisbury-Afshar E, Karnik NS, Joyce C, et al. Publicly available machine learning models for identifying opioid misuse from the clinical notes of hospitalized patients. BMC Med Inform Decis Mak 2020 Apr 29;20(1):79 [FREE Full text] [doi: 10.1186/s12911-020-1099-y] [Medline: 32349766]
135. Lee J, Liu C, Kim JH, Butler A, Shang N, Pang C, et al. Comparative effectiveness of medical concept embedding for feature engineering in phenotyping. JAMIA Open 2021 Apr;4(2):ooab028 [FREE Full text] [doi: 10.1093/jamiaopen/ooab028] [Medline: 34142015]
136. Garcelon N, Neuraz A, Salomon R, Bahi-Buisson N, Amiel J, Picard C, et al. Next generation phenotyping using narrative reports in a rare disease clinical data warehouse. Orphanet J Rare Dis 2018 May 31;13(1):85 [FREE Full text] [doi: 10.1186/s13023-018-0830-6] [Medline: 29855327]
137. Afzal N, Mallipeddi VP, Sohn S, Liu H, Chaudhry R, Scott CG, et al. Natural language processing of clinical notes for identification of critical limb ischemia. Int J Med Inform 2018 Mar;111:83-89 [FREE Full text] [doi: 10.1016/j.ijmedinf.2017.12.024] [Medline: 29425639]
138. Hully M, Lo Barco T, Kaminska A, Barcia G, Cances C, Mignot C, et al. Deep phenotyping unstructured data mining in an extensive pediatric database to unravel a common KCNA2 variant in neurodevelopmental syndromes. Genet Med 2021 May;23(5):968-971 [FREE Full text] [doi: 10.1038/s41436-020-01039-z] [Medline: 33500571]
139. Chen X, Garcelon N, Neuraz A, Billot K, Lelarge M, Bonald T, et al. Phenotypic similarity for rare disease: ciliopathy diagnoses and subtyping. J Biomed Inform 2019 Dec;100:103308 [FREE Full text] [doi: 10.1016/j.jbi.2019.103308] [Medline: 31622800]
140. Bastarache L, Hughey JJ, Goldstein JA, Bastraache JA, Das S, Zaki NC, et al. Improving the phenotype risk score as a scalable approach to identifying patients with Mendelian disease. J Am Med Inform Assoc 2019 Dec 01;26(12):1437-1447 [FREE Full text] [doi: 10.1093/jamia/ocz179] [Medline: 31609419]
141. Stephen R, Boxwala A, Gertman P. Feasibility of using a large clinical data warehouse to automate the selection of diagnostic cohorts. AMIA Annu Symp Proc 2003;2003:1019 [FREE Full text] [Medline: 14728522]
142. Yahi A, Tatonetti NP. A knowledge-based, automated method for phenotyping in the EHR using only clinical pathology reports. AMIA Jt Summits Transl Sci Proc 2015;2015:64-68 [FREE Full text] [Medline: 26306239]
143. Hoffman SR, Vines AI, Halladay JR, Pfaff E, Schiff L, Westreich D, et al. Optimizing research in symptomatic uterine fibroids with development of a computable phenotype for use with electronic health records. Am J Obstet Gynecol 2018 Jun;218(6):610.e1-610.e7 [FREE Full text] [doi: 10.1016/j.ajog.2018.02.002] [Medline: 29432754]
144. Haerian K, Salmasian H, Friedman C. Methods for identifying suicide or suicidal ideation in EHRs. AMIA Annu Symp Proc 2012;2012:1244-1253 [FREE Full text] [Medline: 23304402]
145. Evans RS, Benuzillo J, Horne BD, Lloyd JF, Bradshaw A, Budge D, et al. Automated identification and predictive tools to help identify high-risk heart failure patients: pilot evaluation. J Am Med Inform Assoc 2016 Sep;23(5):872-878. [doi: 10.1093/jamia/ocv197] [Medline: 26911827]

146. Upadhyaya SG, Murphree Jr DH, Ngufor CG, Knight AM, Cronk DJ, Cima RR, et al. Automated diabetes case identification using electronic health record data at a tertiary care facility. Mayo Clin Proc Innov Qual Outcomes 2017 Jul;1(1):100-110 [FREE Full text] [doi: 10.1016/j.mayocpiqo.2017.04.005] [Medline: 30225406]
147. Ahmed A, Thongprayoon C, Pickering BW, Akhoundi A, Wilson G, Pieczkiewicz D, et al. Towards prevention of acute syndromes: electronic identification of at-risk patients during hospital admission. Appl Clin Inform 2014;5(1):58-72 [FREE Full text] [doi: 10.4338/ACI-2013-07-RA-0045] [Medline: 24734124]
148. Redman JS, Natarajan Y, Hou JK, Wang J, Hanif M, Feng H, et al. Accurate identification of fatty liver disease in data warehouse utilizing natural language processing. Dig Dis Sci 2017 Oct;62(10):2713-2718. [doi: 10.1007/s10620-017-4721-9] [Medline: 28861720]
149. Nigwekar SU, Solid CA, Ankers E, Malhotra R, Eggert W, Turchin A, et al. Quantifying a rare disease in administrative data: the example of calciphylaxis. J Gen Intern Med 2014 Aug;29 Suppl 3(Suppl 3):S724-S731 [FREE Full text] [doi: 10.1007/s11606-014-2910-1] [Medline: 25029979]
150. Krebs J, Bittrich M, Dietrich G, Ertl M, Fette G, Kaspar M, et al. Finding needles in the haystack: identifying patients with rare subtype of multiple myeloma supported by a data warehouse and information extraction. Stud Health Technol Inform 2018;253:160-164. [Medline: 30147064]
151. Coquet J, Bozkurt S, Kan KM, Ferrari MK, Blayney DW, Brooks JD, et al. Comparison of orthogonal NLP methods for clinical phenotyping and assessment of bone scan utilization among prostate cancer patients. J Biomed Inform 2019 Jun;94:103184 [FREE Full text] [doi: 10.1016/j.jbi.2019.103184] [Medline: 31014980]
152. Bozkurt S, Paul R, Coquet J, Sun R, Banerjee I, Brooks JD, et al. Phenotyping severity of patient-centered outcomes using clinical notes: a prostate cancer use case. Learn Health Syst 2020 Oct;4(4):e10237 [FREE Full text] [doi: 10.1002/lrh2.10237] [Medline: 33083539]
153. Meystre SM, Heider PM, Kim Y, Aruch DB, Britten CD. Automatic trial eligibility surveillance based on unstructured clinical data. Int J Med Inform 2019 Sep;129:13-19 [FREE Full text] [doi: 10.1016/j.ijmedinf.2019.05.018] [Medline: 31445247]
154. Agarwal V, Podchiyska T, Banda JM, Goel V, Leung TI, Minty EP, et al. Learning statistical models of phenotypes using noisy labeled training data. J Am Med Inform Assoc 2016 Dec;23(6):1166-1173 [FREE Full text] [doi: 10.1093/jamia/ocw028] [Medline: 27174893]
155. Ferté T, Cossin S, Schaeverbeke T, Barnetche T, Jouhet V, Hejblum BP. Automatic phenotyping of electronical health record: PheVis algorithm. J Biomed Inform 2021 May;117:103746 [FREE Full text] [doi: 10.1016/j.jbi.2021.103746] [Medline: 33746080]
156. Chase HS, Radhakrishnan J, Shirazian S, Rao MK, Vawdrey DK. Under-documentation of chronic kidney disease in the electronic health record in outpatients. J Am Med Inform Assoc 2010;17(5):588-594. [doi: 10.1136/jamia.2009.001396] [Medline: 20819869]
157. Kim C, Zhu V, Obeid J, Lenert L. Natural language processing and machine learning algorithm to identify brain MRI reports with acute ischemic stroke. PLoS One 2019;14(2):e0212778 [FREE Full text] [doi: 10.1371/journal.pone.0212778] [Medline: 30818342]
158. Kim JH, Hua M, Whittington RA, Lee J, Liu C, Ta CN, et al. A machine learning approach to identifying delirium from electronic health records. JAMIA Open 2022 Jul;5(2):ooac042 [FREE Full text] [doi: 10.1093/jamiaopen/ooac042] [Medline: 35663114]
159. Zuo X, Li J, Zhao B, Zhou Y, Dong X, Duke J, et al. Normalizing clinical document titles to LOINC document ontology: an initial study. AMIA Annu Symp Proc 2021 Jan 25;2020:1441-1450 [FREE Full text] [Medline: 33936520]
160. Scheurwegs E, Luyckx K, Luyten L, Daelemans W, Van den Bulcke T. Data integration of structured and unstructured sources for assigning clinical codes to patient stays. J Am Med Inform Assoc 2016 Apr;23(e1):e11-e19 [FREE Full text] [doi: 10.1093/jamia/ocv115] [Medline: 26316458]
161. Zhu D, Wu S, Carterette B, Liu H. Using large clinical corpora for query expansion in text-based cohort identification. J Biomed Inform 2014 Jun;49:275-281 [FREE Full text] [doi: 10.1016/j.jbi.2014.03.010] [Medline: 24680983]
162. Tien M, Kashyap R, Wilson GA, Hernandez-Torres V, Jacob AK, Schroeder DR, et al. Retrospective derivation and validation of an automated electronic search algorithm to identify post operative cardiovascular and thromboembolic complications. Appl Clin Inform 2015;6(3):565-576 [FREE Full text] [doi: 10.4338/ACI-2015-03-RA-0026] [Medline: 26448798]
163. Lelong R, Soualmia LF, Grosjean J, Taalba M, Darmoni SJ. Building a semantic health data warehouse in the context of clinical trials: development and usability study. JMIR Med Inform 2019 Dec 20;7(4):e13917 [FREE Full text] [doi: 10.2196/13917] [Medline: 31859675]
164. Pressat-Laffouilhère T, Balayé P, Dahamna B, Lelong R, Billey K, Darmoni SJ, et al. Evaluation of Doc'EDS: a French semantic search tool to query health documents from a clinical data warehouse. BMC Med Inform Decis Mak 2022 Feb 08;22(1):34 [FREE Full text] [doi: 10.1186/s12911-022-01762-4] [Medline: 35135538]
165. Zeng QT, Redd D, Rindflesch T, Nebeker J. Synonym, topic model and predicate-based query expansion for retrieving clinical documents. AMIA Annu Symp Proc 2012;2012:1050-1059 [FREE Full text] [Medline: 23304381]

166. Li M, Lee K, Liu Z, Ma M, Pan Q, Chen R, et al. Applying Bayesian hyperparameter optimization towards accurate and efficient topic modeling in clinical notes. In: Proceedings of the IEEE 9th International Conference on Healthcare Informatics. 2021 Presented at: ICHI '21; August 9-12, 2021; Victoria, BC p. 493-494 URL: https://ieeexplore.ieee.org/document/9565781 [doi: 10.1109/ichi52183.2021.00086]
167. Chen JH, Goldstein MK, Asch SM, Mackey L, Altman RB. Predicting inpatient clinical order patterns with probabilistic topic models vs conventional order sets. J Am Med Inform Assoc 2017 May 01;24(3):472-480 [FREE Full text] [doi: 10.1093/jamia/ocw136] [Medline: 27655861]
168. Afshar M, Joyce C, Dligach D, Sharma B, Kania R, Xie M, et al. Subtypes in patients with opioid misuse: a prognostic enrichment strategy using electronic health record data in hospitalized patients. PLoS One 2019;14(7):e0219717 [FREE Full text] [doi: 10.1371/journal.pone.0219717] [Medline: 31310611]
169. Ling AY, Kurian AW, Caswell-Jin JL, Sledge GW, Shah NH, Tamang SR. Using natural language processing to construct a metastatic breast cancer cohort from linked cancer registry and electronic medical records data. JAMIA Open 2019 Dec;2(4):528-537 [FREE Full text] [doi: 10.1093/jamiaopen/ooz040] [Medline: 32025650]
170. Wu DW, Bernstein JA, Bejerano G. Discovering monogenic patients with a confirmed molecular diagnosis in millions of clinical notes with MonoMiner. Genet Med 2022 Oct;24(10):2091-2102 [FREE Full text] [doi: 10.1016/j.gim.2022.07.008] [Medline: 35976265]
171. Chen CJ, Warikoo N, Chang Y, Chen J, Hsu W. Medical knowledge infused convolutional neural networks for cohort selection in clinical trials. J Am Med Inform Assoc 2019 Nov 01;26(11):1227-1236 [FREE Full text] [doi: 10.1093/jamia/ocz128] [Medline: 31390470]
172. Mutinda FW, Nigo S, Wakamiya S, Aramaki E. Detecting redundancy in electronic medical records using clinical BERT. The Association for Natural Language Processing. 2020. URL: https://www.anlp.jp/proceedings/annual_meeting/2020/pdf_dir/E3-3.pdf [accessed 2023-11-27]
173. Mahajan D, Poddar A, Liang JJ, Lin Y, Prager JM, Suryanarayanan P, et al. Identification of semantically similar sentences in clinical notes: iterative intermediate training using multi-task learning. JMIR Med Inform 2020 Nov 27;8(11):e22508 [FREE Full text] [doi: 10.2196/22508] [Medline: 33245284]
174. Li J, Zhang X, Zhou X. ALBERT-based self-ensemble model with semisupervised learning and data augmentation for clinical semantic textual similarity calculation: algorithm validation study. JMIR Med Inform 2021 Jan 22;9(1):e23086 [FREE Full text] [doi: 10.2196/23086] [Medline: 33480858]
175. Pivovarov R, Elhadad N. A hybrid knowledge-based and data-driven approach to identifying semantically similar concepts. J Biomed Inform 2012 Jun;45(3):471-481 [FREE Full text] [doi: 10.1016/j.jbi.2012.01.002] [Medline: 22289420]
176. Cohen R, Elhadad M, Elhadad N. Redundancy in electronic health record corpora: analysis, impact on text mining performance and mitigation strategies. BMC Bioinformatics 2013 Jan 16;14:10 [FREE Full text] [doi: 10.1186/1471-2105-14-10] [Medline: 23323800]
177. Garcelon N, Neuraz A, Benoit V, Salomon R, Kracker S, Suarez F, et al. Finding patients using similarity measures in a rare diseases-oriented clinical data warehouse: Dr. Warehouse and the needle in the needle stack. J Biomed Inform 2017 Sep;73:51-61 [FREE Full text] [doi: 10.1016/j.jbi.2017.07.016] [Medline: 28754522]
178. Mirzapour M, Abdaoui A, Tchechmedjiev A, Digan W, Bringay S, Jonquet C. French FastContext: a publicly accessible system for detecting negation, temporality and experiencer in French clinical notes. J Biomed Inform 2021 May;117:103733 [FREE Full text] [doi: 10.1016/j.jbi.2021.103733] [Medline: 33737205]
179. Zhou L, Melton GB, Parsons S, Hripcsak G. A temporal constraint structure for extracting temporal information from clinical narrative. J Biomed Inform 2006 Aug;39(4):424-439 [FREE Full text] [doi: 10.1016/j.jbi.2005.07.002] [Medline: 16169282]
180. Klappe ES, van Putten FJ, de Keizer NF, Cornet R. Contextual property detection in Dutch diagnosis descriptions for uncertainty, laterality and temporality. BMC Med Inform Decis Mak 2021 Apr 07;21(1):120 [FREE Full text] [doi: 10.1186/s12911-021-01477-y] [Medline: 33827555]
181. Lin C, Bethard S, Dligach D, Sadeque F, Savova G, Miller TA. Does BERT need domain adaptation for clinical negation detection? J Am Med Inform Assoc 2020 Apr 01;27(4):584-591 [FREE Full text] [doi: 10.1093/jamia/ocaa001] [Medline: 32044989]
182. Garcelon N, Neuraz A, Benoit V, Salomon R, Burgun A. Improving a full-text search engine: the importance of negation detection and family history context to identify cases in a biomedical data warehouse. J Am Med Inform Assoc 2017 May 01;24(3):607-613. [doi: 10.1093/jamia/ocw144] [Medline: 28339516]
183. Cossin S, Jolly M, Larrouture I, Griffier R, Jouhet V. Semi-automatic extraction of abbreviations and their senses from electronic health records. ResearchGate. Preprint posted online July 3, 2023 2021 [FREE Full text]
184. Moon S, Ihrke D, Zeng Y, Liu H. Distinction between medical and non-medical usages of short forms in clinical narratives. AMIA Annu Symp Proc 2017;2017:1302-1311 [FREE Full text] [Medline: 29854199]
185. Peng Y, Yan S, Lu Z. Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets. In: Proceedings of the 18th BioNLP Workshop and Shared Task. 2019 Presented at: BioNLP '19; August 1, 2019; Florence, Italy p. 58-65 URL: https://aclanthology.org/W19-5006.pdf [doi: 10.18653/v1/w19-5006]

186. Peng Y, Yan S, Lu Z. An empirical study of multi-task learning on BERT for biomedical text mining. In: Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing. 2020 Presented at: BioNLP '20; July 9, 2020; Virtual Event p. 205-214 URL: https://aclanthology.org/2020.bionlp-1.22.pdf [doi: 10.18653/v1/2020.bionlp-1.22]
187. Tawfik NS, Spruit MR. Evaluating sentence representations for biomedical text: methods and experimental results. J Biomed Inform 2020 Apr;104:103396. [doi: 10.1016/j.jbi.2020.103396] [Medline: 32147441]
188. Neuraz A, Looten V, Rance B, Daniel N, Garcelon N, Llanos LC, et al. Do you need embeddings trained on a massive specialized corpus for your clinical natural language processing task? Stud Health Technol Inform 2019 Aug 21;264:1558-1559. [doi: 10.3233/SHTI190533] [Medline: 31438230]
189. Dligach D, Afshar M, Miller T. Toward a clinical text encoder: pretraining for clinical natural language processing with applications to substance misuse. J Am Med Inform Assoc 2019 Nov 01;26(11):1272-1278 [FREE Full text] [doi: 10.1093/jamia/ocz072] [Medline: 31233140]
190. Lee YC, Jung S, Kumar A, Shim I, Song M, Kim MS, et al. ICD2Vec: Mathematical representation of diseases. J Biomed Inform 2023 May;141:104361. [doi: 10.1016/j.jbi.2023.104361] [Medline: 37054960]
191. Zhan X, Humbert-Droz M, Mukherjee P, Gevaert O. Structuring clinical text with AI: old versus new natural language processing techniques evaluated on eight common cardiovascular diseases. Patterns (N Y) 2021 Jul 09;2(7):100289 [FREE Full text] [doi: 10.1016/j.patter.2021.100289] [Medline: 34286303]
192. Dubois S, Kale DC, Romano N, Shah N, Jung K. Learning effective representations from clinical Nnotes. arXiv. Preprint posted online May 19, 2017 2017 [FREE Full text] [doi: 10.48550/arXiv.1705.07025]
193. Dynomant E, Lelong R, Dahamna B, Massonnaud C, Kerdelhué G, Grosjean J, et al. Word embedding for the French natural language in health care: comparative study. JMIR Med Inform 2019 Jul 29;7(3):e12310 [FREE Full text] [doi: 10.2196/12310] [Medline: 31359873]
194. Lee D, Jiang X, Yu H. Harmonized representation learning on dynamic EHR graphs. J Biomed Inform 2020 Jun;106:103426 [FREE Full text] [doi: 10.1016/j.jbi.2020.103426] [Medline: 32339747]
195. Roberts K, Si Y, Gandhi A, Bernstam E. A FrameNet for cancer information in clinical narratives: schema and annotation. In: Proceedings of the 11th International Conference on Language Resources and Evaluation. 2018 Presented at: LREC '18; July 15-18, 2018; Miyazaki, Japan p. 272-279 URL: https://aclanthology.org/L18-1041.pdf
196. Van Vleck TT, Stein DM, Stetson PD, Johnson SB. Assessing data relevance for automated generation of a clinical summary. AMIA Annu Symp Proc 2007 Oct 11;2007:761-765 [FREE Full text] [Medline: 18693939]
197. Escudié JB, Jannot AS, Zapletal E, Cohen S, Malamut G, Burgun A, et al. Reviewing 741 patients records in two hours with FASTVISU. AMIA Annu Symp Proc 2015;2015:553-559 [FREE Full text] [Medline: 26958189]
198. Feller DJ, Zucker J, Don't Walk OB, Srikishan B, Martinez R, Evans H, et al. Towards the inference of social and behavioral determinants of sexual health: development of a gold-standard corpus with semi-supervised learning. AMIA Annu Symp Proc 2018;2018:422-429 [FREE Full text] [Medline: 30815082]
199. Loda S, Krebs J, Danhof S, Schreder M, Solimando AG, Strifler S, et al. Exploration of artificial intelligence use with ARIES in multiple myeloma research. J Clin Med 2019 Jul 09;8(7):999 [FREE Full text] [doi: 10.3390/jcm8070999] [Medline: 31324026]
200. Song H, Gu Y, Leroy G, Donovan FM, Galgiani JN. Integrating automated biomedical lexicon creation for valley fever diagnosis. In: Proceedings of the 2021 IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies. 2021 Presented at: CHASE '21; December 16-17, 2021; Washington, DC p. 111-112 URL: https://ieeexplore.ieee.org/document/9697921 [doi: 10.1109/chase52844.2021.00021]
201. Patrick J, Li M. High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. J Am Med Inform Assoc 2010 Oct;17(5):524-527 [FREE Full text] [doi: 10.1136/jamia.2010.003939] [Medline: 20819856]
202. Patrick JD, Nguyen DH, Wang Y, Li M. A knowledge discovery and reuse pipeline for information extraction in clinical notes. J Am Med Inform Assoc 2011;18(5):574-579 [FREE Full text] [doi: 10.1136/amiajnl-2011-000302] [Medline: 21737844]
203. Chen L, Gu Y, Ji X, Lou C, Sun Z, Li H, et al. Clinical trial cohort selection based on multi-level rule-based natural language processing system. J Am Med Inform Assoc 2019 Nov 01;26(11):1218-1226 [FREE Full text] [doi: 10.1093/jamia/ocz109] [Medline: 31300825]
204. Solt I, Tikk D, Gál V, Kardkovács ZT. Semantic classification of diseases in discharge summaries using a context-aware rule-based classifier. J Am Med Inform Assoc 2009;16(4):580-584 [FREE Full text] [doi: 10.1197/jamia.M3087] [Medline: 19390101]
205. Uzuner Ö. Recognizing obesity and comorbidities in sparse data. J Am Med Inform Assoc 2009;16(4):561-570 [FREE Full text] [doi: 10.1197/jamia.M3115] [Medline: 19390096]
206. Yang X, Lyu T, Li Q, Lee C, Bian J, Hogan WR, et al. A study of deep learning methods for de-identification of clinical notes in cross-institute settings. BMC Med Inform Decis Mak 2019 Dec 05;19(Suppl 5):232 [FREE Full text] [doi: 10.1186/s12911-019-0935-4] [Medline: 31801524]

207. Ferrández O, South BR, Shen S, Friedlin FJ, Samore MH, Meystre SM. BoB, a best-of-breed automated text de-identification system for VHA clinical documents. J Am Med Inform Assoc 2013 Jan 01;20(1):77-83 [FREE Full text] [doi: 10.1136/amiajnl-2012-001020] [Medline: 22947391]
208. Woo H, Kim K, Cha K, Lee J, Mun H, Cho SJ, et al. Application of efficient data cleaning using text clustering for semistructured medical reports to large-scale stool examination reports: methodology study. J Med Internet Res 2019 Jan 08;21(1):e10013 [FREE Full text] [doi: 10.2196/10013] [Medline: 30622098]
209. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Proceedings of the 31st Conference on Neural Information Processing Systems. 2017 Presented at: NIPS '17; December 4-9, 2017; Long Beach, CA p. 1-11 URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
210. Uzuner Ö, Solti I, Cadag E. Extracting medication information from clinical text. J Am Med Inform Assoc 2010;17(5):514-518 [FREE Full text] [doi: 10.1136/jamia.2010.003947] [Medline: 20819854]

Abbreviations

BERT: Bidirectional Encoder Representations from Transformers
BiLSTM-CRF: bidirectional long short-term memory–conditional random field
CDW: clinical data warehouse
CNN: convolutional neural network
CRF: conditional random field
cTAKES: clinical Text Analysis and Knowledge Extraction System
EHR: electronic health record
ELMo: embeddings from language models
GloVe: global vectors for word representation
i2b2: Informatics for Integrating Biology & the Bedside
ICD-9: International Classification of Diseases, Ninth Revision
LSTM: long short-term memory
MedLEE: Medical Language Extraction and Encoding System
n2c2: National NLP Clinical Challenges
NCBO: National Center for Biomedical Ontology
NER: named entity recognition
NLP: natural language processing
PHI: protected health information
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
SNOMED-CT: Systematized Nomenclature of Medicine–Clinical Terms
SVM: support vector machine
TF-IDF: term frequency–inverse document frequency
UMLS: Unified Medical Language System

Edited by C Lovis; submitted 05.09.22; peer-reviewed by M Behzadifar, MF Kabir, B Hoyt; comments to author 17.11.22; revised version received 16.01.23; accepted 07.09.23; published 15.12.23.

Please cite as:
Bazoge A, Morin E, Daille B, Gourraud PA
Applying Natural Language Processing to Textual Data From Clinical Data Warehouses: Systematic Review
JMIR Med Inform 2023;11:e42477
URL: https://medinform.jmir.org/2023/1/e42477
doi:10.2196/42477
PMID:38100200

A Call to Reconsider a Nationwide Electronic Health Record System: Correcting the Failures of the National Program for IT

James Seymour Morris, BA

School of Clinical Medicine, University of Cambridge, Addenbrooke's Hospital NHS Foundation Trust, Cambridge, United Kingdom

Corresponding Author:
James Seymour Morris, BA

Related Article:

This is a corrected version. See correction statement: https://medinform.jmir.org/2024/1/e56050

Abstract

The National Programme for IT (NPfIT) was launched in 2005 to implement 7 nationwide IT services across the National Health Service (NHS). Despite the success of many of these designated "deliverables," the establishment of a single nationwide electronic health record (EHR) system never fully materialized. As a result, NHS medical records are now stored using a diverse array of alternate EHR systems, which frequently restricts health care practitioners from accessing extensive portions of their patients' notes. This not only limits their ability to make well-informed clinical decisions but also impacts the quality of care they are able to provide. This article assesses the medical, economic, and bureaucratic implications of an NHS-wide EHR system. Additionally, it explores how the shortcomings of the NPfIT should be addressed when attempting to introduce such a system in the future.

(JMIR Med Inform 2023;11:e53112) doi:10.2196/53112

KEYWORDS

electronic health record; EHR; medical record linkage; health information interoperability; health information management; health information systems; information systems; interoperability; health records; medical records; national

"The Biggest IT Failure Ever Seen" [1]

In April 2005, the NHS Connecting for Health Agency (CFH) was established to implement the National Programme for IT (NPfIT). Its goal was to propel NHS England into the 21st century by creating 7 nationwide IT services, including a secure NHSmail system and an electronic prescription service. Although many of these services have now been incorporated into everyday practice, the program's primary objective was never realized: to establish a single nationwide database of electronic health records (EHRs). Reluctance from several trusts to transition from paper-based records, along with concerns regarding delayed financial returns and data privacy, led to the NPfIT being branded as "the biggest IT failure ever seen" [1], and it was ultimately discontinued. Now, on the decennary of the program's dissolution, it is increasingly apparent that a nationwide EHR service could represent the panacea for the inexcusable inefficiencies within today's NHS.

The Status Quo

Instead of a single nationwide EHR system, NHS trusts currently have the option to choose from 40 different approved suppliers [2]. Due to economic competition, these suppliers offer limited cross-compatibility. NHS staff are well aware of the restrictions this places on the quality of health care. Oftentimes, we cannot access notes from previous admissions due to their being recorded using an alternate EHR system, nor do we have the luxury of previous laboratory results and radiographs with which to compare recent investigations. During the 12 months following April 2017, 3.9 million patients in NHS England attended more than one trust as either an inpatient or outpatient, or in accident and emergency department [3]. On over 11 million occasions, patients sought care at a trust that used a different EHR system than their previous attendance, implying that for 9.1% of all acute presentations [3], clinicians were left uninformed about their patient's recent medical history.

Following a hospital admission, general practitioners (GPs) are left with little more than a brief discharge summary from which they assimilate weeks' worth of inpatient notes. To access the results of inpatient investigations, they must first overcome administrative hurdles, and even then, detailed notes are seldom made available to GPs. Although notes can be transferred, most often via GP2GP [4] when switching primary care providers, data are notoriously "lost in translation" between different EHR systems. Furthermore, administrative delays completely preclude the transfer of EHR data in acute health care settings. Such obstacles quickly leave NHS staff feeling not unlike the protagonist of Kafka's 1926 novel, "The Castle" [5]: so disillusioned by the excessive complexity of bureaucracy that

they settle for the status quo, ceasing to seek the information they presume to be unattainable.

Creating a single national database would not only improve the quality of care for contemporary patients but also benefit generations to come. Clinical research would achieve unprecedented statistical power if physicians were granted access to the full cohort of patients registered with NHS GPs—comprising over 62 million people in England alone [6]. Although national research databases are available through the Data Access Request Service, they provide incomplete information of limited applicability and are are hindered by a thick layer of bureaucratic red tape.

Finally, the burden of manually reporting notifiable diseases and maternal deaths could be entirely avoided by automating the process within a nationwide EHR system, thereby improving both the efficiency and fidelity of MBRRACE-UK (Mothers and Babies: Reducing Risk through Audit and Confidential Enquiries) and UKHSA (United Kingdom Health Security Agency) data.

Lessons From NPfIT

The status quo of NHS record-keeping is indisputably Kafkaesque, but if we are to successfully establish a nationwide database the second time around, we must first address the NPfIT's shortcomings. The Committee of Public Accounts published a damning report outlining the reasons for the program's failure; of these, poor financial returns received the lion's share of the blame. The £9.8 billion investment had yielded just £3.7 billion out of a forecasted £10.7 billion by March 2012 [7] (£1.00 GBP = $1.27 USD at the time of writing); however, this was largely due to the reluctancy of many NHS trusts to switch to the nationwide system, with only 22 of 220 trusts joining the integrated network. Without full participation, the government not only had a weak position in negotiating licensing costs [7], but it was also unable to reap the aforementioned benefits to efficiency and service quality that come with a fully unified system. Therefore, the reported £3.7 billion is likely to represent but a fraction of the potential savings under the final system. The licensing costs for 5 years were originally quoted at £3.1 million per trust [7] (then, £682 million in total), suggesting that the majority of the costs up until the program's dissolution were one-time investments in establishing the new EHR system, in addition to the £31.5 million squandered on prematurely terminating existing contracts with Fujitsu [7]. The costs quoted are, therefore, likely to diminish alongside exponentially increasing benefits as the system is implemented throughout the NHS.

There have been concerns that granting a single provider with a monopoly on EHRs in the NHS could result in drastic price hikes and limit service quality, with critics quoting the old adage that "competition breeds innovation." This is true, but only with regard to innovations that maximize profits. By incentivizing co-operation rather than competition between developers (for example, through establishing open source software development), beneficial features would be accreted rather than credentialed as intellectual property, which Carson aptly criticizes as a "toll on the free transfer of information" [8]. In addition, there is evidence to suggest that countries with stricter market entry regulations, and thus, restricted economic competition, do not suffer from reduced quality in public or private goods and services [9].

To motivate universal participation throughout the NHS, the concerns of its employees must also be addressed; namely, the unintuitive user interface and limited functionality of all EHR systems during the early days of their development. This caused many trusts to stick with paper-based records and, as of May 2022, 27 NHS trusts had still not transitioned to digital records [10]. Although the most popular EHR systems are still considered to have suboptimal usability [11], the principal barrier today will be convincing trusts to switch from one EHR to the nationwide system, rather than phasing out paper-based systems, as was the original issue. Parenthetically, EHR system usability varies significantly between different developers [11], and as such, implementing a single nationwide system should also help to curtail geographical health care inequalities by providing a more uniform service.

The NPfIT was further criticized for failing to provide training for NHS staff on how to use the novel EHR systems, which did not initially allow clinicians to view laboratory results and other investigations within the same interface. The former issue is easily resolved by offering training to staff who are unfamiliar with the new EHR system using a platform like e-Learning for Healthcare. Conversely, the latter emphasizes the need for the nationwide EHR system to be a "one-stop shop" for all patient data. The interface should combine both primary and secondary care notes, along with investigations and imaging within a single interface, a concept which, in primary care settings, has already been demonstrated to improve cost-effectiveness, efficiency, and patient satisfaction [12].

Using a single platform to store the comprehensive medical records of an entire population does, nonetheless, bring with it a risk of massive data breaches. Accordingly, the NPfIT and its predecessors were heavily criticized for failing to establish adequate safeguards against such a scenario in which patient data would be vulnerable to surveillance by governmental and clandestine organisations. The unsettling imagery this provokes—that of an Orwellian panopticon endlessly monitoring the public's medical records—led Privacy International to present the NPfIT with a Big Brother Award for the "most appalling project" of 2004. Since then, the public has grown far more supportive of a nationwide system, with only 9.6% of participants from a 2013 study [13], and 12% of participants from a 2015 study [14], opposing the nationwide EHR system. This shift in attitudes has coincided with increasing safeguards as well as the introduction of mandatory annual training on data protection for all NHS staff. Nevertheless, a unified system should still demand further safeguards, such as limiting the access of the most sensitive sections of the record, for example, genitourinary medicine, fertility, and psychiatry, to those health care professionals directly involved in a given patient's care.

Looking to the Future

Switching to a single nationwide system will, of course, demand a massive investment of time and resources; however, should

it be attempted, it has the potential to slash NHS expenditure for decades to come. Policymakers will have to garner the support of most, if not all, 215 NHS trusts, if the system is to prove financially viable and avoid falling victim to the same shortcomings as its predecessor. Once a contract is negotiated, a firm date should be set, by which time all NHS trusts are required to phase out their current EHR system in favor of the nationwide network. Such a date should lie beyond the expiry of all contemporary contracts to circumvent the expenses associated with their premature termination, an oversight which ultimately contributed to the downfall of the NPfIT. Fortuitously, the majority of the concerns raised against the program, both by clinicians and the general public, have since been overcome. Nonetheless, e-learning modules should be introduced during the interregnum to allow health care staff to familiarize themselves with the new platform and to emphasize the importance of high-quality data entry prior to its formal introduction. The EHR system itself will have to be refined at regular intervals if it is to continue meeting the ever-evolving needs of its various stakeholders. It is essential, therefore, that a public health body be established to represent the needs of each NHS trust and to continuously re-evaluate them as part of the NHS commissioning cycle. When these are no longer being met, such an organization should present these shortcomings to developers and advocate for specific enhancements to be made to the software.

By providing NHS physicians and approved researchers with a cradle-to-grave record of all NHS patients, the nationwide EHR system would allow clinical decisions to become far better informed and, as a result, the quality of health care should also drastically improve. With consideration for the myriad benefits outlined hitherto, the Department of Health and Social Care is urged not to abandon the nationwide EHR system as Kafka did "The Castle" [5], but to see it through until clinicians and patients alike may reap these benefits for years to come.

Acknowledgments

I would like to thank my close personal friend, Pauline Gronczewska, for our engaging debates regarding the inherent inefficiencies of NHS EHR systems and, ultimately, the inspiration for this manuscript.

Conflicts of Interest

None declared.

References

1. Syal R. Abandoned NHS IT system has cost £10bn so far. The Guardian. 2013 Sep 18. URL: www.theguardian.com/society/2013/sep/18/nhs-records-system-10bn [accessed 2023-07-06]
2. NHS England accredited supplier lists. NHS England. URL: www.england.nhs.uk/hssf/supplier-lists/#enterprise-wide-electronic-patient-records-systems-for-acute-community-and-mental-health-hospitals [accessed 2023-07-06]
3. Warren LR, Clarke J, Arora S, Darzi A. Improving data sharing between acute hospitals in England: an overview of health record system distribution and retrospective observational analysis of inter-hospital transitions of care. BMJ Open 2019 Dec 5;9(12):e031637. [doi: 10.1136/bmjopen-2019-031637] [Medline: 31806611]
4. GP2GP. NHS Digital. 2023 Dec 6. URL: https://digital.nhs.uk/services/gp2gp [accessed 2023-12-24]
5. Kafka F. Das Schloss [Book in German], 1st edition: Kurt Wolff Verlag; 1926.
6. Patients registered at a GP practice, July 2023. NHS Digital. 2023. URL: https://digital.nhs.uk/data-and-information/publications/statistical/patients-registered-at-a-gp-practice/july-2023 [accessed 2023-07-06]
7. The dismantled national programme for IT in the NHS. Nineteenth report of session 2013–14. House of Commons Committee of Public Accounts. 2013 Sep 18. URL: https://publications.parliament.uk/pa/cm201314/cmselect/cmpubacc/294/294.pdf [accessed 2023-07-06]
8. Carson KA. Organization Theory: A Libertarian Perspective, 1st edition: Booksurge LLC. URL: https://kevinacarson.org/pdf/ot.pdf [accessed 2023-12-18]
9. Djankov S, La Porta R, Lopez-de-Silanes F, Shleifer A. The regulation of entry. Q J Econ 2002 Feb 1;117(1):1-37 [FREE Full text] [doi: 10.1162/003355302753399436]
10. Carding N, Harding R. Revealed: the 27 trusts still without an electronic patient record. Health Service Journal. 2022 May 25. URL: www.hsj.co.uk/technology-and-innovation/revealed-the-27-trusts-still-without-an-electronic-patient-record/7032511.article [accessed 2023-12-18]
11. Bloom BM, Pott J, Thomas S, Gaunt DR, Hughes TC. Usability of electronic health record systems in UK EDs. Emerg Med J 2021 Jun;38(6):410-415. [doi: 10.1136/emermed-2020-210401] [Medline: 33658268]
12. Murtagh S, McCombe G, Broughan J, et al. Integrating primary and secondary care to enhance chronic disease management: a scoping review. Int J Integr Care 2021 Feb 9;21(1):4. [doi: 10.5334/ijic.5508] [Medline: 33613136]
13. Luchenski SA, Reed JE, Marston C, Papoutsi C, Majeed A, Bell D. Patient and public views on electronic health records and their uses in the United Kingdom: cross-sectional survey. J Med Internet Res 2013 Aug 23;15(8):e160. [doi: 10.2196/jmir.2701] [Medline: 23975239]

14. Papoutsi C, Reed JE, Marston C, Lewis R, Majeed A, Bell D. Patient and public views about the security and privacy of electronic health records (Ehrs) in the UK: results from a mixed methods study. BMC Med Inform Decis Mak 2015 Oct 14;15:86. [doi: 10.1186/s12911-015-0202-2] [Medline: 26466787]

Abbreviations

EHR: electronic health record
GP: general practitioner
MBRRACE-UK: Mothers and Babies: Reducing Risk through Audit and Confidential Enquiries
NHS: National Health Service
NPfIT: National Programme for IT
UKHSA: United Kingdom Health Security Agency

Edited by C Lovis; submitted 26.09.23; peer-reviewed by A Yazdanian, C Thies, L Roa, U Hübner; revised version received 30.11.23; accepted 02.12.23; published 28.12.23.

Please cite as:
Morris JS
A Call to Reconsider a Nationwide Electronic Health Record System: Correcting the Failures of the National Program for IT
JMIR Med Inform 2023;11:e53112
URL: https://medinform.jmir.org/2023/1/e53112
doi:10.2196/53112

Viewpoint

Developing a Capsule Clinic—A 24-Hour Institution for Improving Primary Health Care Accessibility: Evidence From China

Dongliang Li[1*], PhD; Rujia Zhang[2*], BPA; Chun Chen[2*], PhD; Yunyun Huang[2], BPA; Xiaoyi Wang[2], BPA; Qingren Yang[2], BPA; Xuebo Zhu[2], PhD; Xiangyang Zhang[3], MPA; Mo Hao[1*], PhD; Liming Shui[4*], MPH

[1]School of Public Health, Fudan University, Shanghai, China

[2]School of Public Health and Management, Wenzhou Medical University, Wenzhou, China

[3]Engineering Research Center of Intelligent Medicine (2016E10011), The First Affiliated Hospital of Wenzhou Medical University, Wenzhou, China

[4]Yinzhou District Health Bureau, Ningbo, China

[*]these authors contributed equally

Corresponding Author:
Liming Shui, MPH
Yinzhou District Health Bureau
No 1221, Bachelor Road, Shounan Street, Yinzhou District
Ningbo, 315199
China
Phone: 86 13967820698
Email: 707761065@qq.com

Abstract

Telehealth is an effective combination of medical service and intelligent technology. It can improve the problem of remote access to medical care. However, an imbalance in the allocation of health resources still occurs. People spend more time and money to access higher-quality services, which results in inequitable access to primary health care (PHC). At the same time, patients' usage of telehealth services is limited by the equipment and their own knowledge, and the PHC service suffers from low usage efficiency and lack of service supply. Therefore, improving PHC accessibility is crucial to narrowing the global health care coverage gap and maintaining health equity. In recent years, China has explored several new approaches to improve PHC accessibility. One such approach is the capsule clinic, an emerging institution that represents an upgraded version of the internet hospital. In coordination with the United Nations, the Yinzhou district of Ningbo city in Zhejiang, China, has been testing this new model since 2020. As of October 2022, the number of applications in Ningbo was 15, and the number of users reached 12,219. Unlike internet hospitals, the entire process—from diagnosis to prescription services—can be completed at the capsule clinic. The 24-hour telehealth service could also solve transportation problems and save time for users. Big data analysis can accurately identify regional populations' PHC service needs and improve efficiency in health resource allocation. The user-friendly, low-cost, and easily accessible telehealth model is of great significance. Installation of capsule clinics would improve PHC accessibility and resolve the uneven distribution of health resources to promote health equity.

(JMIR Med Inform 2023;11:e41212) doi:10.2196/41212

KEYWORDS

primary health care; accessibility; capsule clinic; 24-hour clinic; big-data; China; United Nations; internet clinic

Background

Combining intelligent technology with health care has become increasingly popular [1,2]. Telehealth is the remote provision of clinical health care and health administration services via information and telecommunication technologies [3]. In exploring suitable methods to meet people's increasing demands, countries are establishing intelligent medical models in telehealth to provide convenient health services [4] and solve geographical, temporal, and economic problems of accessibility to health care. Doctors use telehealth to transmit digital imaging, conduct video consultations, and make medical diagnoses. Telehealth began in the 20th century [5], with the advent of television, and has developed rapidly in the 21st century along with advancements in technology. Most people currently have access to basic devices, such as mobile phones and computers, which can be used to obtain telehealth services. With improved accessibility of health care through telehealth, individuals in rural areas and busy urban areas can often connect more easily with a provider. In the United States and the United Kingdom

[6], telehealth is a popular trend. According to the National Health Service Long Term Plan in the United Kingdom, "digitally enabled care will go mainstream" [7].

Compared with other countries, China is a late starter in the field of telehealth. In recent years, China has been committed to applying internet technology to medical services. The internet hospital is a rapidly developing, new telehealth model that is gaining wide popularity [8]. The establishment of internet hospitals could help some citizens overcome temporal and geographical barriers to accessing traditional medical services [9]. However, the issue of accessibility to primary health care (PHC) services has not been completely resolved. Owing to urban-rural differences and unbalanced allocation of regional health human resources [10], the phenomenon of inequality in health resource allocation still exists [11]. Furthermore, as the population continues to age [12] and the number of chronic diseases grows [13], the demand for health services increases every year. Researchers have found that accessibility of PHC services has been challenged by insufficient funds [14], limited-service locations [15], poverty-stricken areas, and inconvenient transportation [16-18]. The use of PHC is inefficient because people rely on doctors in higher-tier hospitals [19]. Although the internet hospital is a promising public health tool because it could significantly increase access to health care for medically underserved populations, some obstacles remain. For example, people with low digital information literacy, especially older people and children, cannot easily use internet health care. Low digital literacy remains the key barrier to accessing intelligent health systems [20,21]. In order to use telehealth, such as internet hospitals, mobile devices have high requirements [22,23]. Inadequate communication facilities hinder telehealth services [24,25] for many people in remote areas, and many older people experience barriers to using smartphones. Internet hospitals must improve applications for older adults. In addition, patients also face the problem of waiting for drug delivery after receiving remote consultation services via the internet hospital, which could cause a loss of time [26,27]. Moreover, big data is not being used well in the allocation of drug resources; the advantages of big data analysis of population characteristics and rational allocation of drug resources cannot be realized in internet hospitals.

In many cases, although internet hospitals address the problem of unbalanced distribution of health resources, the shortcomings mentioned above must be addressed to provide access to PHC services. In the near future, the application of telehealth will not only increase convenience and access to health care but also reduce costs. Telehealth services should also be more user-friendly and implement low-cost devices to facilitate remote health care services. Intelligent medical service systems should be better integrated with the traditional medical models, and internet hospitals must be upgraded to meet the needs of all members of society.

Capsule Clinics in China

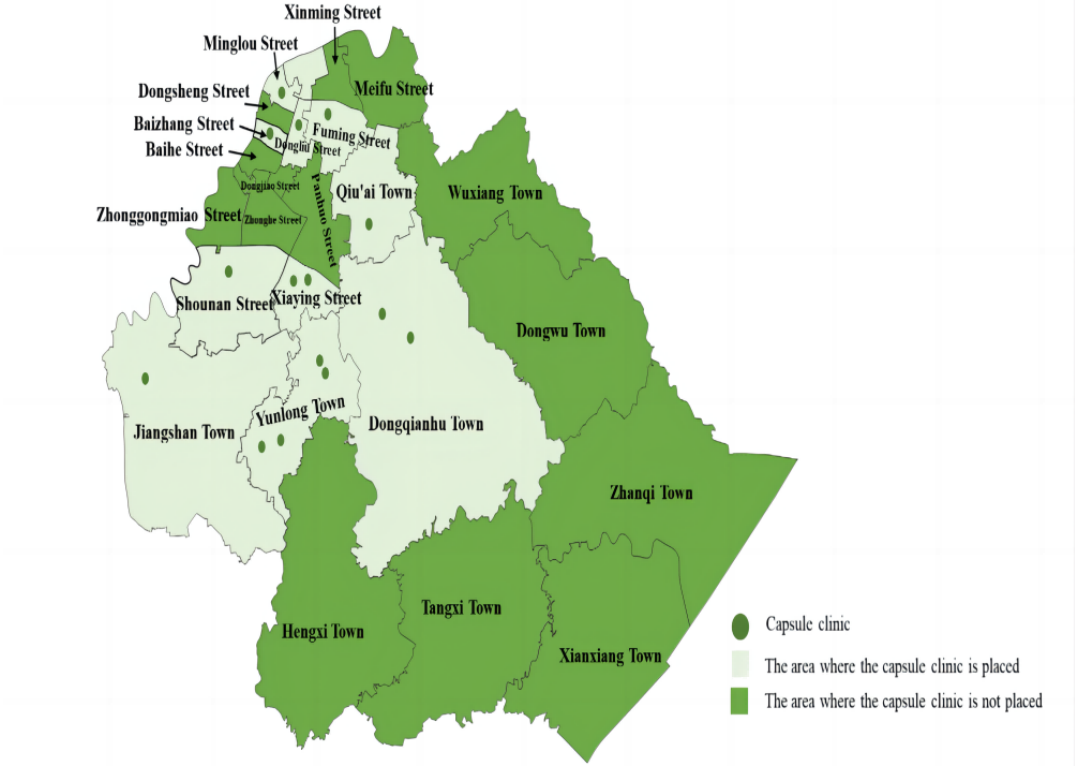
The accessibility of PHC services in China is usually considered in terms of geographical accessibility [28-30], economic accessibility, and temporal accessibility [31,32]. To that end, we propose the development of the Internet Hospital 2.0 in the era of big data, in the context of Ningbo, a city in China with an estimated population of over 9 million individuals [33]. Ningbo residents face challenges in accessing PHC services. They remain disappointed with the inconvenient primary care model and are anxious about the time and economic costs of daily medical visits, particularly during the COVID-19 pandemic. Ningbo is a coastal city that includes many islands and other remote areas. People living in remote villages or islands experience greater challenges in accessing and paying for PHC services.

In an attempt to combine the use of available public health data with the provision of easy access to health care, and toward the achievement of universal health coverage, the World Health Organization implemented a pilot project for developing a higher-quality and efficient medical and health service e-system. The Yinzhou district in Ningbo is one of the pilot locations. Yinzhou has created a specialized health service model in the Chinese context—the capsule clinic, an upgraded version of the internet hospital. The capsule clinic (20 m^2 in area) is a novel type of health service facility that began operating in 2020. It incorporates traditional medical models, including diagnosis and prescription. Capsule clinic services comprise 3 parts: health examination, consultation, and an intelligent pharmacy.

The special clinic, which relies on Zhejiang province's excellence in digital reform and Yinzhou's superior health care resources, is an emerging form of medical treatment facility that can provide residents with more convenient PHC services than those currently available. One advantage of advanced intelligent technology–based models is that they increase geographical accessibility to PHC. Capsule clinics are placed within communities to enable residents' convenient access to PHC without leaving the community. The capsule clinic offers some important advantages, such as 24-hour services, comprehensive medical services that include the entire process from diagnosis to prescription services, and efficient allocation of health resource services. The capsule clinic overcomes some of the administrative shortcomings of the internet hospital.

In recent years, Zhejiang has attached great importance to the combination of intelligent technology and medical services, allowing grassroots residents to enjoy higher-quality and inclusive PHC services. Figure 1 shows that as of October 2022, there were 15 applications for capsule clinics in Yinzhou (Figure 1). According to China's "one village, one health care room" principle, Yinzhou plans to establish over 190 capsule clinics across communities or villages and expand throughout China.

Figure 1. Distribution of capsule clinics in Yinzhou.



The Function of the Capsule Clinic

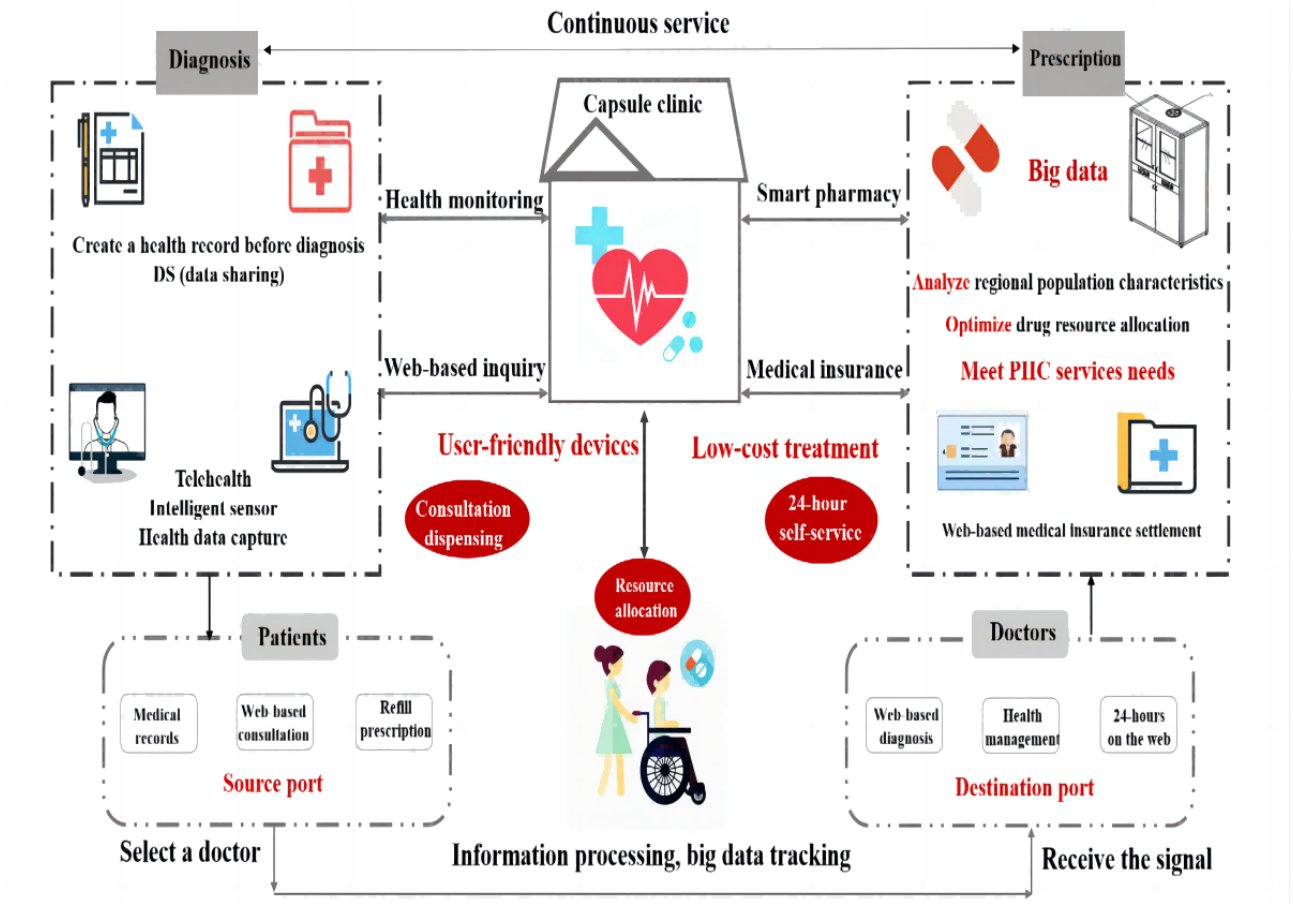
Figure 2 shows that several critical hardware and software components are included in the capsule clinic to ensure the necessary functions to meet the medical demands of residents living nearby. The capsule clinic has continuous medical services, including diagnosis and prescription, which fit the traditional medical model. Doctors can communicate with patients for remote health care consultations around the clock. Patients can access services on their own any time they need them and even gain access to essential medicine. In addition, patients can use the telehealth service of capsule clinics to consult doctors from higher-tier hospitals; this feature provides an opportunity for PHC service resources to be equitably deployed, especially for people in remote areas.

The capsule clinic, a product of the digital reform, has developed rapidly and is devoted to improving geographic accessibility to PHC. For each community, the capsule clinic will be conveniently located within 15 minutes of the community's health service center, which is responsible for the full scope of the capsule clinic. At nearby capsule clinics, people located in remote areas (mountainous areas and islands) can access health resources equal to those available to urban residents. This measure is intended to benefit the public and solve issues of not only geographic accessibility but also temporal accessibility; with the capsule clinic, people do not have to miss work or travel long distances for medical appointments.

Although some internet hospitals and telehealth centers are also available 24 hours a day, capsule clinics are more helpful to residents because they provide comprehensive services. In the past, patients received their medications within a few days after receiving a web-based diagnosis. The capsule clinic takes advantage of intelligent technology to help residents save time in receiving their medications. After completing a remote consultation with the doctor, patients can easily access the capsule clinic's intelligent pharmacy to pick up their medication. They can also refill prescriptions from offline doctors. Moreover, if further examination or tests are needed, patients can visit the higher-tier hospitals within the period determined by the capsule clinic doctors. If hospitalization is required, treatment can be transferred to the hospital, as determined by the capsule clinic doctors. The capsule system embodies the advantages of an integrated health care delivery system. It facilitates convenient access to health care and makes good use of PHC resources.

Big data analysis and artificial intelligence can assist doctors in a variety of ways, such as by detecting lesions and improving diagnostic efficiency. They also play a role in improving medical services and easing constraints on medical resources. With the deep integration of digital technology and medical measures, as well as comprehensive improvements in patients' health literacy, diversified medical and health service models can provide convenience for doctors and patients, thus improving the efficiency of diagnosis and treatment. Big data allows for systematic analysis of the relevant characteristics and medical needs of people in each region and the subsequent reasonable allocation of drug resources. For instance, in a community with a wide distribution of older or chronic patient populations, the capsule clinic would be equipped with more drugs to treat geriatric issues and chronic diseases. One benefit of a big data–based system is the reduction in medical processing time and distance, which significantly improves the efficiency and quality of the medical care provided and facilitates optimum resource allocation.

Figure 2. Main functions of the capsule clinic. PHC: primary health care.



Comparing the Capsule Clinic With the Traditional Hospital and Internet Hospital

Overview

The capsule clinic has the physical appearance and essential equipment of a traditional medical institution. It also includes 24-hour remote consultation, integrated consultation and pharmacy services, optimal health resource allocation, and health management functions. The clinic has the strong support and oversight of a traditional hospital, and offers web-based services for relatively simple problems that do not necessitate a visit to the hospital.

Digital technology will soon be deeply integrated with medical treatment and will provide important support for medical professionals' diagnosis and treatment decisions. However, the internet hospital model [34,35] still has much room for improvement, for example, in terms of medical insurance and the allocation of medical resources. With the support of big data technology, capsule clinics analyze the regional group characteristics, realize the reasonable allocation of essential drug resources, and meet the health needs of different populations. Even more importantly, capsule clinics may alleviate shortages in human resources chronically faced by primary-level medical institutions. The capsule clinic allows community residents to conveniently purchase medication 24 hours a day, taking pressure off in-person facilities. Remote consultation can also reduce pressures on grassroots medical staff and, thus, improve work efficiency. Figure 3 provides comparison of 3 medical models.

In order to understand the user experience of the capsule clinic during the implementation process, we conducted the research during November and December 2021. The research team visited communities and villages where capsule clinics were established, such as Haichuang Community, Lijia Village in Yunlong Town, and Dongfu Community in Qianhu Street, to conduct qualitative interviews with users (Table 1). Qualitative interviews collected basic user characteristics, home addresses, user acceptance of the capsule clinic, and feelings about using it. The interviews revealed that internet hospitals are different from capsule clinics, which have physical clinics and equipment, and rely on medical resources from offline hospitals. Users can not only experience the same diagnosis and prescription services in capsule clinics as in offline hospitals, but also use intelligent equipment that improves the accessibility of PHC services. As capsule clinics are physical entities, patients view them as more reliable than intangible internet hospitals. Moreover, this physicality makes them more user-friendly and greatly reduces use disparities caused by the digital divide [36-38].

Figure 3. Comparison of the 3 medical models. PHC: primary health care.

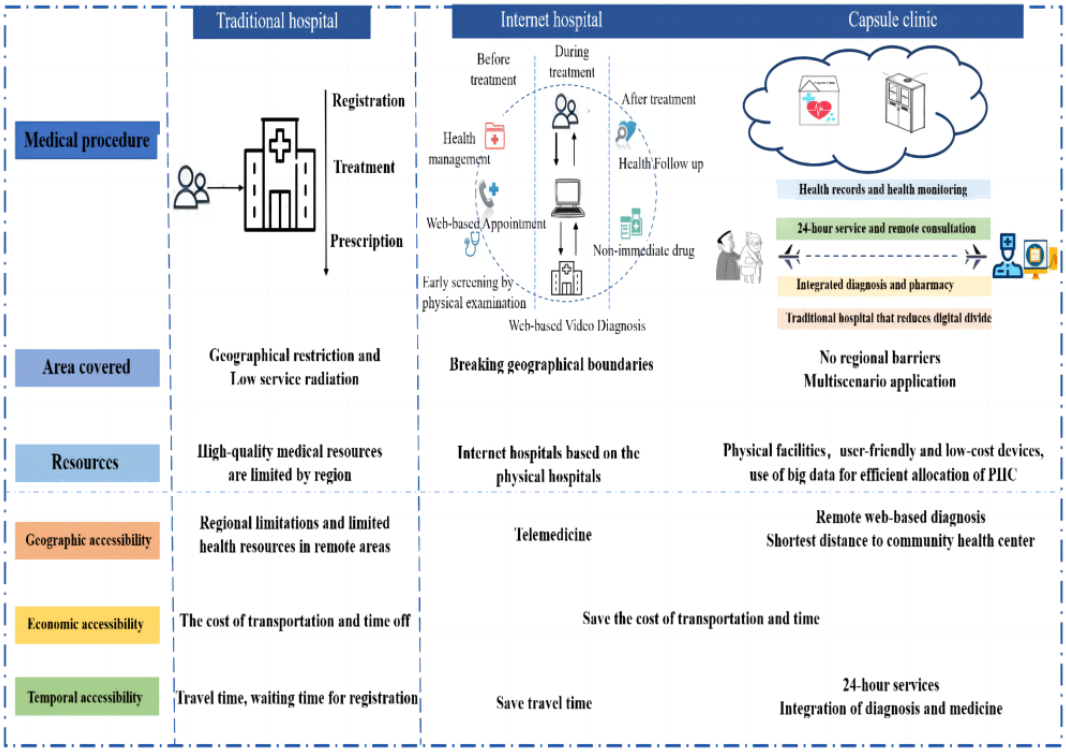


Table 1. Qualitative interviews of the user experience of capsule clinics.

Interviewee	Personal information				Interview content
	Age (years)	Gender	Address of capsule clinic	Job	
Interviewee 1	45	Female	Haichuang Community Health Service Station, Yinzhou District, Ningbo	Employee of a state-owned enterprise	“It is time-saving for me to go to the clinic after work. It is the same as the local community hospital and I can find a doctor remotely through the capsule clinic whenever I need.”
Interviewee 2	28	Male	Hefeng Community Health Service Station, Hefeng Creative Plaza, Yinzhou District, Ningbo	Employee of a foreign company	“I've used capsule clinics accessible by walking. Compared with the Internet hospital, the capsule clinic can give me the feeling of being treated in a physical hospital. I could complete the procedure of diagnosis and treatment and did not have to wait for drugs.”
Interviewee 3	50	Male	Shanglijia Village Committee, Yunlong Town, Yinzhou District, Ningbo	Factory worker	“The capsule clinic is in our village, it's like a convenience store, and I can go in and take my blood pressure when I'm walking by. I come here to take my blood pressure and use the screening program every day.”
Interviewee 4	68	Female	Zhongxing Community Health Service Station, Dongliu Street, Yinzhou District, Ningbo	Retiree	“I haven't been to the capsule clinic but I've heard about it because I'm too old to use these smart devices.”
Interviewee 5	38	Female	Party Mass Service Center of Junrui Community, Xiaying Street, Yinzhou District, Ningbo	Full-time housewife	“The capsule clinic is closer, which is more convenient than the community hospital. The common drugs that my children and older adults need can be dispensed here under the consultation of the doctor.”
Interviewee 6	35	Male	Guanying Village Committee, Yunlong Town, Yinzhou District, Ningbo	Private entrepreneur	“I like the 24-hour service of the capsule clinic, which is convenient and fast.”

Operation Process of the Capsule Clinic

The patient arrives at the capsule clinic and enters. They scan their identification or insurance card to gain access to the web-based site, where they can review their medical records, refill prescriptions, and schedule a remote appointment with a doctor. Figure 4 illustrates this flowchart of capsule clinic health care services for patients. After receiving the signal that a patient needs a consultation, a doctor from a higher-tier medical institution provides web-based medical services, including face-to-face consultation via a remote video system. Figure 5 illustrates this flowchart of capsule clinic health care services for doctors. As noted, patient prescriptions are generally filled immediately after the web-based consultation. The big data system collects the characteristics of the population in the area where the capsule clinic is located, along with past medication habits. It configures the drug resources in the intelligent pharmacy in accordance with the provided information and the doctor's advice. The available drug resources are configured to meet the needs of more than 90% of the population. Patient prescriptions are provided by an intelligent pharmacy that focuses on the community residents' demands and their physical conditions. Furthermore, if residents only want to refill medications at the capsule clinic, they can receive the same prescription with the help of web-based doctors.

Unlike the internet hospital, the capsule clinic not only provides consultation services but also supports patients' medication needs with the intelligent medicine cabinet under the approval of the offline physical clinic. This measure provides great convenience for patients with long-term drug demands due to chronic diseases. When patients complete a physical hospital visit and have qualified for web-based prescription refills, they can pick up their next refill at the capsule clinic. If patients are identified as needing further examination or hospitalization during the remote consultation, the doctor will schedule an appointment for them at the parent hospital.

The number of capsule clinic users reached 12,219 by October 2022 (Table 2). Since 2020, the number of users has shown a gradual upward trend. Most of these patients are aged between 18 and 60 years; older adults have problems using the capsule clinics owing to the digital divide. Regarding the choice of drug purchased at the capsule clinic, 94.09% of patients buy Rx (Receptor X) medication, whereas only 4.49% choose over-the-counter medication, and 1.42% visit the capsule clinic simply to use the health monitoring program. This shows that most patients use both diagnosis and treatment and medicine dispensing functions in capsule clinics. In terms of the total cost of related drugs, more than 80% of patients spent less than 300 Chinese Yuan (US $43.20) per visit. Perhaps most of the PHC needs of nearby residents can be met by the capsule clinics, where people can avail of basic diagnostic services and buy essential medicines.

Various data from the past 3 years show that the number of capsule clinics in use is increasing year by year. Residents who live nearby are curious about the new medical institutions and may try the health monitoring program when walking past. Users are very interested in the 24-hour service provided by the clinic. In addition, the low number of users may have resulted from inadequate publicity; therefore, many people, especially older adults, are afraid to try this novel medical model. Although older adults still experience obstacles to using the new intelligent medical model, the use of capsule clinics is on the rise owing to the rapid development of internet technology, the expansion of internet access facility coverage, and the facilitation of using facilities. The gap between patients' "willingness to use" and "ability to use" capsule clinics may influence the digital divide.

Thus far, based on the geographical distribution mentioned above, the capsule clinics are distributed in each residential location, which satisfies the residents' desire to avail of PHC services close to home and at any time. Compared with the distance that the residents previously had to travel to reach community hospitals, the geographic accessibility of PHC services has been largely addressed by the capsule clinics. Furthermore, the issue of economic accessibility has also been addressed as people with jobs can avoid the cost of absence for medical treatment and the cost of transportation. With the help of big data, the capsule clinic's smart pharmacy could estimate and allocate drugs in accordance with residents' medication characteristics, meet residents' medication needs, and achieve good coverage of PHC services.

Figure 4. Flowchart of capsule clinic health care services for patients.

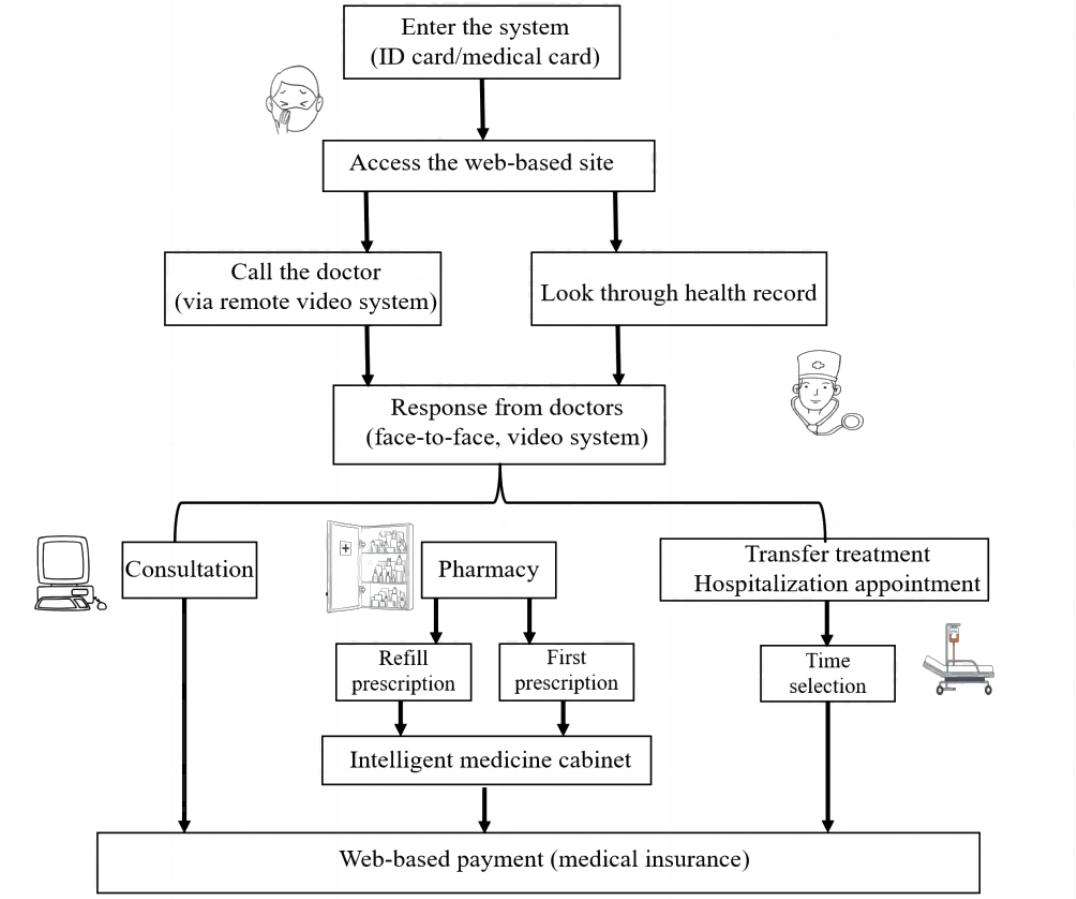


Figure 5. Flowchart of capsule clinic health care services for doctors.

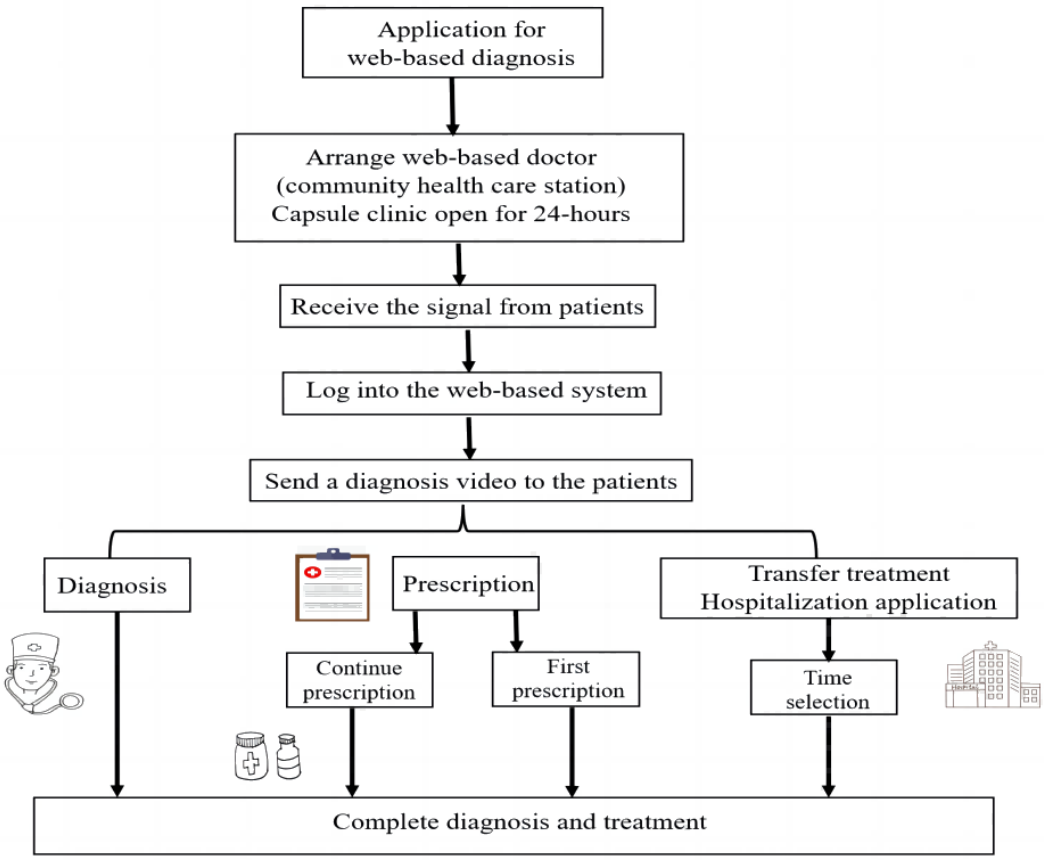


Table 2. Characteristics of patients of capsule clinics from 2020 to 2022.

Characteristics	Patients, n (%)			
	2020	2021	2022	Total
Total users	2392 (19.58)	4595 (37.60)	5232 (42.82)	12,219 (100)
Age (years)				
≤18	430 (17.98)	879 (19.13)	925 (17.68)	2234 (18.28)
18-40	874 (36.54)	1923 (41.85)	1740 (33.25)	4537 (37.13)
40-60	739 (30.89)	1168 (25.42)	1650 (31.54)	3557 (29.11)
≥60	349 (14.59)	625 (13.60)	917 (17.53)	1891 (15.48)
Gender				
Male	839 (35.08)	1531 (33.32)	1982 (37.88)	4352 (35.62)
Female	1553 (64.92)	3064 (66.68)	3250 (62.12)	7867 (64.38)
Medicine category				
Rx[a]	2212 (92.48)	4297 (93.52)	4989 (95.35)	11,498 (94.09)
OTC[b]	84 (3.51)	250 (5.44)	214 (4.09)	548 (4.49)
No medicine[c]	96 (4.01)	48 (1.04)	29 (0.56)	173 (1.42)
Total drug cost range (Chinese Yuan[d])				
0-100	957 (40.01)	1742 (37.91)	2048 (39.15)	4747 (38.85)
100-200	872 (36.46)	1492 (32.47)	1664 (31.80)	4028 (32.97)
200-300	251 (10.49)	544 (11.84)	481 (9.19)	1276 (10.44)
>300	312 (13.04)	817 (17.78)	1039 (19.86)	2168 (17.74)

[a]Rx: Receptor X.

[b]OTC: over the counter.

[c]No medicine: patients do not obtain medicine from the capsule clinic and only use the health monitoring system services, such as measuring blood pressure, height, weight, and other basic items.

[d]1 Chinese Yuan=US $0.14.

Routine Management and Operation

In the context of the United Nations (UN) pilot program, since the capsule clinic concept will be implemented throughout China, its routine management should be considered. The physical medical institutions (community health care stations) that have obtained the "Medical Institution Practicing License" can apply to establish a capsule clinic. The institution and application report should be submitted together to the local health administration department. Doctors from the parent institution provide remote consultations for the associated capsule clinic. Additionally, traditional hospitals (community health care centers), doctors, and the local health administration department are jointly responsible for the routine management. It is important to encourage video surveillance at capsule clinics and to incorporate procedures that ensure the traceability of drugs dispensed automatically. The physical health care center is primarily responsible for not only the unified management of capsule clinics set up by subordinate branches but also the safety of the medical services provided and the quality of the drugs dispensed at the capsule clinic. Drugs dispensed by the intelligent medicine cabinet at the capsule clinic should adhere to the wholesale drug distribution enterprises' unified procurement guidelines.

Conclusions and Limitations

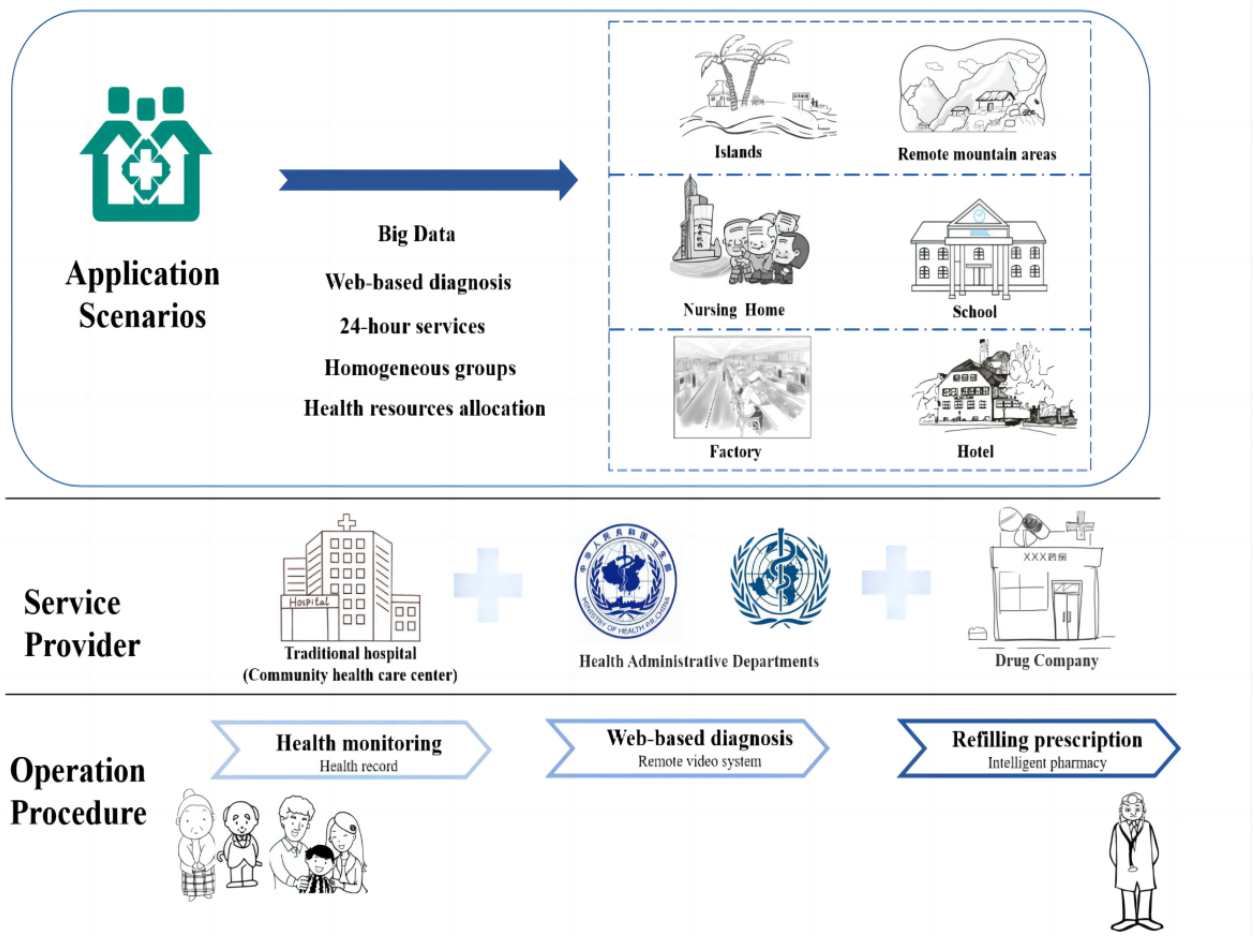
With the emergence of capsule clinics, China has a new medical model that can alleviate PHC accessibility problems. Capsule clinics are currently being promoted in every township, village, and community in Ningbo, with the goal of ensuring that every community resident has access to PHC services at their doorstep. The clinics make it convenient for people in remote areas to achieve higher-quality and equitable health resources. In the future, the UN and China will reach consensus on promoting capsule clinics nationwide and become the global template for PHC service delivery. Capsule clinics in remote areas can improve the geographic accessibility for people who lack PHC resources. Capsules can also work well in cities through the use of big data. In many cases and situations, capsule clinics can provide people with the timely PHC services they need, thus reducing financial losses.

Capsule clinics are aimed at people with both common diseases and chronic diseases, and they provide web-based follow-up services, medication disbursement, web-based medical insurance settlement, and more. With the goal of enhanced PHC accessibility, the UN and China intend to place capsule clinics in schools, nursing homes, suburbs, islands, remote mountain

areas, and other modern communities. Establishing capsules in these areas can have important effects on many populations. Figure 6 shows that the capsule clinics can be used in the future as follows:

1. Capsule clinics can be placed in remote mountainous areas and islands to increase residents' access to PHC services. In many cases, low geographic accessibility is a barrier to residents' access to effective health resources. The capsule clinic can help solve the problem. People in remote areas can receive basic health care services through telehealth. The advantages of capsule clinics, especially in terms of access to essential medicines, are greatly highlighted for these populations.
2. Capsule clinics can use big data to effectively allocate resources, accurately identify regional population characteristics, allocate appropriate PHC resources, and increase investments in the resources needed to meet the various needs of the people. As such, capsule clinics can be placed in homogeneous, crowded places such as schools and nursing homes. As a special place in the community, the school takes the safety of students very seriously, and an on-site capsule clinic can help ensure it.
3. Communities should take advantage of capsule clinics' 24-hour services. Capsule clinics could be installed in crowded locations with nighttime activity to prevent accidents and, at the same time, offer people timely treatment or access to medication. For example, factories and hotels are key places where things go wrong; an on-site capsule would allow people to seek timely medical help. Doctors are always available digitally to provide reliable advice for seekers.

Figure 6. Future application scenarios planned.



Of course, the capsule clinic has some limitations. First, data security supervision and drug use safety must be strengthened. Second, the publicity for capsule clinics is not sufficient. People seem to be unaware of this new model or have a wait-and-see attitude. Influenced by traditional health care concepts, numerous patients are more likely to use physical hospitals and are unwilling to use the capsule clinic; consequently, capsule clinic usage is still low. Third, capsule clinics in China are currently limited to urban communities in Ningbo; the single location application and scenario reflect the lack of promotion efforts.

Although the capsule clinic still has some challenges and limitations, it is meaningful in terms of developing convenient and less costly intelligent PHC treatments. People who live in remote areas such as villages and poor mountainous areas far from health care centers face barriers to accessing PHC. Telehealth needs to be enhanced in terms of user experience, and the equipment should be more user-friendly and universal. People are more likely to use smart medical treatment to improve the quality of health services on the basis of the familiar medical treatment mode. Many patients are likely to be driven into poverty by the indirect economic burden of disease. Meanwhile, with higher living standards and enhanced

conceptions of health, more and more people are demanding better health products and higher-quality health services. If we are to reach our goal of improving universal health coverage, we must commit to investing in and scaling up proven solutions. Thus, promoting capsule clinics is highly relevant. The intelligent medical industry is still in its infancy, and there is great room for improvement and sufficient potential demand. In the future, the capsule clinic may help solve fundamental imbalances in the distribution of medical resources and contradictions among the growing health care needs of the population.

Countries need to make more and smarter investments in foundational health systems, with an emphasis on PHC, essential services, and marginalized populations. We must make great efforts to ensure equal access to public health services for all.

Acknowledgments

We would like to thank Yinzhou District Health Bureau, Ningbo, for their assistance and support in the field research. This study was sponsored by The National Natural Science Foundation of China (72274141); the Zhejiang Provincial Natural Science Foundation (LY22G030006); the World Health Organization, which collaborated on a pilot project to build a high-quality and efficient health service system (GJ2-2021- WHOPO-C2); and the General Research Project of Education Department (Y202147813). The sponsors were not involved in the design and conduct of the study; the collection, management, analysis, and interpretation of data; or the preparation, review, and approval of the manuscript.

Authors' Contributions

DLL, RJZ, CC, MH, and LMS conceived and designed the study. YYH, XYW, XYZ, and QRY participated in field research. CC, RJZ, YYH, and XBZ wrote the original draft. DLL, RJZ, LMS, and XYZ substantively revised the manuscript. All authors have read and approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Tian S, Yang W, Grange JML, Wang P, Huang W, Ye Z. Smart healthcare: making medical care more intelligent. Glob Health J 2019 Sep;3(3):62-65. [doi: 10.1016/j.glohj.2019.07.001]
2. Cáceres C, Rosário J, Amaya D. Proposal of a smart hospital based on Internet of Things (IoT) concept. 2019 Presented at: SaMBa 2018: Processing and Analysis of Biomedical Information; September 20, 2018; Granada. [doi: 10.1007/978-3-030-13835-6_11]
3. Dorsey ER, Topol EJ. State of telehealth. N Engl J Med 2016 Jul 14;375(2):154-161. [doi: 10.1056/NEJMra1601705] [Medline: 27410924]
4. Dorsey ER, Topol EJ. Telemedicine 2020 and the next decade. Lancet 2020 Mar 14;395(10227):859. [doi: 10.1016/S0140-6736(20)30424-4] [Medline: 32171399]
5. Makhni MC, Riew GJ, Sumathipala MG. Telemedicine in orthopaedic surgery: challenges and opportunities. J Bone Joint Surg Am 2020 Jul 01;102(13):1109-1115. [doi: 10.2106/JBJS.20.00452] [Medline: 32618908]
6. Fischer SH, Ray KN, Mehrotra A, Bloom EL, Uscher-Pines L. Prevalence and characteristics of telehealth utilization in the United States. JAMA Netw Open 2020 Oct 01;3(10):e2022302 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.22302] [Medline: 33104208]
7. Digitally enabled care to go ‘mainstream’ in the next decade, Long Term Plan promises. Digital Health. URL: https://tinyurl.com/3rk4xse3 [accessed 2022-12-13]
8. Han Y, Lie RK, Guo R. The internet hospital as a telehealth model in China: systematic search and content analysis. J Med Internet Res 2020 Jul 29;22(7):e17995 [FREE Full text] [doi: 10.2196/17995] [Medline: 32723721]
9. Zhang J, Lu Q, Shi L. The influence of telemedicine on capacity development in public primary hospitals in China: a scoping review. Clinical eHealth 2022 Dec;5:91-99. [doi: 10.1016/j.ceh.2022.10.001]
10. Dong E, Liu S, Chen M, Wang H, Chen L, Xu T, et al. Differences in regional distribution and inequality in health-resource allocation at hospital and primary health centre levels: a longitudinal study in Shanghai, China. BMJ Open 2020 Jul 19;10(7):e035635 [FREE Full text] [doi: 10.1136/bmjopen-2019-035635] [Medline: 32690509]
11. Wang Y, Li Y, Qin S, Kong Y, Yu X, Guo K, et al. The disequilibrium in the distribution of the primary health workforce among eight economic regions and between rural and urban areas in China. Int J Equity Health 2020 Feb 26;19(1):28 [FREE Full text] [doi: 10.1186/s12939-020-1139-3] [Medline: 32102655]
12. Han Y, He Y, Lyu J, Yu C, Bian M, Lee L. Aging in China: perspectives on public health. Glob Health J 2020 Mar;4(1):11-17. [doi: 10.1016/j.glohj.2020.01.002]
13. Zhao Y, Atun R, Oldenburg B, McPake B, Tang S, Mercer SW, et al. Physical multimorbidity, health service use, and catastrophic health expenditure by socioeconomic groups in China: an analysis of population-based panel data. Lancet Glob Health 2020 Jun;8(6):e840-e849 [FREE Full text] [doi: 10.1016/S2214-109X(20)30127-3] [Medline: 32446349]

14. Johar M, Soewondo P, Pujisubekti R, Satrio HK, Adji A. Inequality in access to health care, health insurance and the role of supply factors. Soc Sci Med 2018 Sep;213:134-145. [doi: 10.1016/j.socscimed.2018.07.044] [Medline: 30077959]
15. Song Y, Tan Y, Song Y, Wu P, Cheng JC, Kim MJ, et al. Spatial and temporal variations of spatial population accessibility to public hospitals: a case study of rural–urban comparison. GIsci Remote Sens 2018 Mar 15;55(5):718-744. [doi: 10.1080/15481603.2018.1446713]
16. Swere KMR. Challenges hindering the accessibility of Tanzania's health service: a literature review. Int J Econ Finance 2016 Jul 20;8(8):242. [doi: 10.5539/ijef.v8n8p242]
17. Benevenuto RG, Azevedo ICC, Caulfield B. Assessing the spatial burden in health care accessibility of low-income families in rural Northeast Brazil. J Transp Health 2019 Sep;14:100595. [doi: 10.1016/j.jth.2019.100595]
18. Cao W, Shakya P, Karmacharya B, Xu DR, Hao Y, Lai Y. Equity of geographical access to public health facilities in Nepal. BMJ Glob Health 2021 Oct 27;6(10):e006786 [FREE Full text] [doi: 10.1136/bmjgh-2021-006786] [Medline: 34706879]
19. Zhang Y, Wang Q, Jiang T, Wang J. Equity and efficiency of primary health care resource allocation in mainland China. Int J Equity Health 2018 Sep 12;17(1):140 [FREE Full text] [doi: 10.1186/s12939-018-0851-8] [Medline: 30208890]
20. Lee T, Lee B, Lee-Geiller S. The effects of information literacy on trust in government websites: evidence from an online experiment. Int J Inf Manage 2020 Jun;52:102098. [doi: 10.1016/j.ijinfomgt.2020.102098]
21. Kruse C, Fohn J, Wilson N, Nunez Patlan E, Zipp S, Mileski M. Utilization barriers and medical outcomes commensurate with the use of telehealth among older adults: systematic review. JMIR Med Inform 2020 Aug 12;8(8):e20359 [FREE Full text] [doi: 10.2196/20359] [Medline: 32784177]
22. Maassen O, Fritsch S, Gantner J, Deffge S, Kunze J, Marx G, et al. Future mobile device usage, requirements, and expectations of physicians in German university hospitals: web-based survey. J Med Internet Res 2020 Dec 21;22(12):e23955 [FREE Full text] [doi: 10.2196/23955] [Medline: 33346735]
23. Al-Turjman F, Nawaz MH, Ulusar UD. Intelligence in the Internet of Medical Things era: A systematic review of current and future trends. Comput Commun 2020 Jan;150:644-660. [doi: 10.1016/j.comcom.2019.12.030]
24. Cortelyou-Ward K, Atkins DN, Noblin A, Rotarius T, White P, Carey C. Navigating the digital divide: barriers to telehealth in rural areas. J Health Care Poor Underserved 2020;31(4):1546-1556. [doi: 10.1353/hpu.2020.0116] [Medline: 33416736]
25. Almathami HKY, Win KT, Vlahu-Gjorgievska E. Barriers and facilitators that influence telemedicine-based, real-time, online consultation at patients' homes: systematic literature review. J Med Internet Res 2020 Feb 20;22(2):e16407 [FREE Full text] [doi: 10.2196/16407] [Medline: 32130131]
26. Usak M, Kubiatko M, Shabbir MS, Viktorovna Dudnik O, Jermsittiparsert K, Rajabion L. Health care service delivery based on the Internet of things: a systematic and comprehensive study. Int J Commun Syst 2019 Sep 13;33(2):e4179. [doi: 10.1002/dac.4179]
27. Javaid M, Khan IH. Internet of Things (IoT) enabled healthcare helps to take the challenges of COVID-19 Pandemic. J Oral Biol Craniofac Res 2021;11(2):209-214 [FREE Full text] [doi: 10.1016/j.jobcr.2021.01.015] [Medline: 33665069]
28. Yin C, He Q, Liu Y, Chen W, Gao Y. Inequality of public health and its role in spatial accessibility to medical facilities in China. Appl Geogr 2018 Mar;92:50-62. [doi: 10.1016/j.apgeog.2018.01.011]
29. Wang X, Yang H, Duan Z, Pan J. Spatial accessibility of primary health care in China: A case study in Sichuan Province. Soc Sci Med 2018 Jul;209:14-24. [doi: 10.1016/j.socscimed.2018.05.023] [Medline: 29778934]
30. Wang Y, Wu X, Yin M, Jin L. Patient capability: justice and grassroots healthcare delivery in China. Dev World Bioeth 2022 Sep;22(3):170-178. [doi: 10.1111/dewb.12328] [Medline: 34342130]
31. Zhang W, Ung COL, Lin G, Liu J, Li W, Hu H, et al. Factors contributing to patients' preferences for primary health care institutions in China: a qualitative study. Front Public Health 2020;8:414 [FREE Full text] [doi: 10.3389/fpubh.2020.00414] [Medline: 33014959]
32. Bitton A, Ratcliffe HL, Veillard JH, Kress DH, Barkley S, Kimball M, et al. Primary health care as a foundation for strengthening health systems in low- and middle-income countries. J Gen Intern Med 2017 May;32(5):566-571 [FREE Full text] [doi: 10.1007/s11606-016-3898-5] [Medline: 27943038]
33. The Seventh National Census of the Main Data Bulletin, Ningbo, Zhejiang. Ningbo Municipal Bureau of Statistics. URL: http://tjj.ningbo.gov.cn/art/2021/5/17/art_1229042825_58913572.html [accessed 2021-05-17]
34. Zhi L, Yin P, Ren J, Wei G, Zhou J, Wu J, et al. Running an internet hospital in China: perspective based on a case study. J Med Internet Res 2021 Sep 16;23(9):e18307 [FREE Full text] [doi: 10.2196/18307] [Medline: 34342267]
35. Xie X, Zhou W, Lin L, Fan S, Lin F, Wang L, et al. Internet hospitals in China: cross-sectional survey. J Med Internet Res 2017 Jul 04;19(7):e239 [FREE Full text] [doi: 10.2196/jmir.7854] [Medline: 28676472]
36. Hong YA, Zhou Z, Fang Y, Shi L. The digital divide and health disparities in China: evidence from a national survey and policy implications. J Med Internet Res 2017 Sep 11;19(9):e317 [FREE Full text] [doi: 10.2196/jmir.7786] [Medline: 28893724]
37. Yang Y, Zhang X, Lee PK. Improving the effectiveness of online healthcare platforms: an empirical study with multi-period patient-doctor consultation data. Int J Prod Econ 2019 Jan;207:70-80. [doi: 10.1016/j.ijpe.2018.11.009]
38. Li C, Zhang E, Han J. Adoption of online follow-up service by patients: an empirical study based on the elaboration likelihood model. Comput Hum Behav 2021 Jan;114:106581. [doi: 10.1016/j.chb.2020.106581]

Abbreviations

PHC: primary health care
Rx: Receptor X
UN: United Nations

Edited by C Lovis, J Hefner; submitted 19.07.22; peer-reviewed by W Dong, S Wei; comments to author 05.10.22; revised version received 18.11.22; accepted 08.12.22; published 09.01.23.

Please cite as:
Li D, Zhang R, Chen C, Huang Y, Wang X, Yang Q, Zhu X, Zhang X, Hao M, Shui L
Developing a Capsule Clinic—A 24-Hour Institution for Improving Primary Health Care Accessibility: Evidence From China
JMIR Med Inform 2023;11:e41212
URL: https://medinform.jmir.org/2023/1/e41212
doi:10.2196/41212
PMID:36622737

Viewpoint

One Digital Health Intervention for Monitoring Human and Animal Welfare in Smart Cities: Viewpoint and Use Case

Arriel Benis[1,2,3*], PhD; Mostafa Haghi[4,5,6*], PhD; Thomas M Deserno[5,6], PhD; Oscar Tamburis[2,3,7], PhD

[1]Department of Digital Medical Technologies, Holon Institute of Technology, Holon, Israel

[2]Working Group "One Digital Health", European Federation for Medical Informatics (EFMI), Le Mont-sur-Lausanne, Switzerland

[3]Working Group "One Digital Health", International Medical Informatics Association (IMIA), Chene-Bourg, Geneva, Switzerland

[4]Ubiquitous Computing Laboratory, Department of Computer Science, HTWG Konstanz – University of Applied Sciences, Konstanz, Germany

[5]Peter L. Reichertz Institute for Medical Informatics of TU Braunschweig and Hannover Medical School, Braunschweig, Germany

[6]Working Group "Accident & Emergency Informatics", International Medical Informatics Association (IMIA), Chene-Bourg, Geneva, Switzerland

[7]Institute of Biostructures and Bioimaging, National Research Council, Naples, Italy

[*]these authors contributed equally

Corresponding Author:
Arriel Benis, PhD
Department of Digital Medical Technologies
Holon Institute of Technology
52 Golomb Street
Holon, 5810201
Israel
Phone: 972 03 5026892
Email: arrielb@hit.ac.il

Abstract

Smart cities and digital public health are closely related. Managing digital transformation in urbanization and living spaces is challenging. It is critical to prioritize the emotional and physical health and well-being of humans and their animals in the dynamic and ever-changing environment they share. Human-animal bonds are continuous as they live together or share urban spaces and have a mutual impact on each other's health as well as the surrounding environment. In addition, sensors embedded in the Internet of Things are everywhere in smart cities. They monitor events and provide appropriate responses. In this regard, accident and emergency informatics (A&EI) offers tools to identify and manage overtime hazards and disruptive events. Such manifold focuses fit with One Digital Health (ODH), which aims to transform health ecosystems with digital technology by proposing a comprehensive framework to manage data and support health-oriented policies. We showed and discussed how, by developing the concept of ODH intervention, the ODH framework can support the comprehensive monitoring and analysis of daily life events of humans and animals in technologically integrated environments such as smart homes and smart cities. We developed an ODH intervention use case in which A&EI mechanisms run in the background. The ODH framework structures the related data collection and analysis to enhance the understanding of human, animal, and environment interactions and associated outcomes. The use case looks at the daily journey of Tracy, a healthy woman aged 27 years, and her dog Mego. Using medical Internet of Things, their activities are continuously monitored and analyzed to prevent or manage any kind of health-related abnormality. We reported and commented on an ODH intervention as an example of a real-life ODH implementation. We gave the reader examples of a "how-to" analysis of Tracy and Mego's daily life activities as part of a timely implementation of the ODH framework. For each activity, relationships to the ODH dimensions were scored, and relevant technical fields were evaluated in light of the Findable, Accessible, Interoperable, and Reusable principles. This "how-to" can be used as a template for further analyses. An ODH intervention is based on Findable, Accessible, Interoperable, and Reusable data and real-time processing for global health monitoring, emergency management, and research. The data should be collected and analyzed continuously in a spatial-temporal domain to detect changes in behavior, trends, and emergencies. The information periodically gathered should serve human, animal, and environmental health interventions by providing professionals and caregivers with inputs and "how-to's" to improve health, welfare, and risk prevention at the individual and population levels. Thus, ODH complementarily combined with A&EI is meant to enhance policies and systems and modernize emergency management.

(JMIR Med Inform 2023;11:e43871) doi:10.2196/43871

KEYWORDS

One Health; Digital Health; One Digital Health; accident and emergency informatics; eHealth; informatics; medicine; veterinary medicine; environmental monitoring; education; patient engagement; citizen science; data science; pets; human-animal bond; intervention; ambulatory monitoring; health monitoring; Internet of Things; smart environment; mobile phone

Introduction

Background

The current data-driven society pushes people to seek and ask for an always-increasing number of highly personalized services [1]. Citizens of smart cities use ubiquitous and mobile technologies and expect to obtain what they need before asking for it. Thus, smart cities are models of urbanization development wherein technologies are actively deployed to enhance the quality of life (QoL) of both humans and animals [2].

Accordingly, personal data such as location, interests (eg, queries on search engines, reads on social media, answers to surveys, and purchases), contacts (eg, calls and social media connections), and activities (eg, neighborhood meetings and personal announcements) are shared continuously [3]. Concerning health-related issues, citizens, as patients, demand proactive suggestions from a modern and evolving public health sector to improve self-care and lower the number of critical events. To this end, their digitized health records encompass the entire process, from the physician's consultation (in person or remote) to purchasing prescribed drugs (in person or on the web). People use web-based services to check their laboratory tests or reports of certified physical training. Furthermore, they may use applications for sleep monitoring that are provided by a health care management organization [4-8].

Smart cities and (digital) public health share aspects related to a healthy lifestyle. The Internet of Things (IoT; we will use *IoT* as an encompassing term including the Internet of Medical Things [IoMT] and the Internet of Animal Health Things [IoAHT]) monitors several subtopics of public health: environmental conditions, electromagnetic radiation, health conditions, fitness activities, food quality, emotions, and accidents [9-11]. Therefore, we need signals, data, and information produced by sensors in wearables or mobile apps or existing on social media networks to run health interventions [10] and cope with emergencies [12]. Accordingly, this has been called to expand the capabilities of health records. In smart cities, IoT devices also trace and track animals. Such applications reflect social and cultural norms, the safety and wellness of animals (eg, pet activity and feeding trackers [13]), and the collective composed of humans and animals [14]. In such communities, large amounts of data are produced continuously and exchanged wirelessly [15]. Its purposes range from home automation to trip management and QoL. These sensors range from microelectromechanical systems to advanced medical devices.

It is worth noting that, during the first waves of the COVID-19 pandemic, an increasing rate of shared information was observed. The need arose to timely deliver health care for both humans and animals as specific clusters of customers belonging to the same ecosystem [5,9]. Valid data collection was needed [16,17] that fit the essential elements of smart (healthy) cities [18,19]. For instance, McConnell et al [20] analyzed the influence of owning animals on stress. Other work has assessed the impact of the surrounding ecosystem on humans' and animals' QoL [21,22].

A practical setup for the entire data management process comprises aspects of generation, collection, transmission, storage, extraction, analysis, reporting, and decision-making as codified according to the principles of accident and emergency informatics (A&EI), with the dual purpose of preventing harm and supporting decisions [23]. Furthermore, aspects such as education, citizen engagement, and the large vision of human nature are notable elements of the One Digital Health (ODH) framework.

The ODH Framework to Set Up an ODH Intervention

The *ODH framework* develops around 2 so-called keys [24]. On the one hand, One Health looks at monitoring and assessing environmental hazard interactions and their impacts on health and biodiversity [25]. In contrast, Digital Health stands as the mature deployment of currently available technology to improve individuals' health and care [26]. The framework also includes 3 perspectives, 5 dimensions, and a technological ring.

As ODH proposes a new holistic, data-driven approach encompassing human, animal, and environmental health and welfare, a solid common ground with this working scenario emerges. Table 1 adjusts the common aspects of smart cities and digital public health using the ODH framework.

Health professionals traditionally define an intervention as the whole process, starting by defining a protocol; implementing it; systematically evaluating its effects; and then using the results to define, for example, a new therapeutic strategy or a new policy. According to the Digital Health perspective, we refer to *ODH Intervention* as the use of digital and mobile technologies for specific initiatives addressing human, animal, and environmental system needs. In other words, an ODH intervention is a real-world implementation of the theoretical ODH framework, which relies on Findable, Accessible, Interoperable, and Reusable (FAIR) data [27-29]. In this regard, we recognize an appropriate alignment with the *Classification of Digital Health Interventions v1.0* proposed by the World Health Organization (WHO), which deals with the principles of implementing and delivering formal or informal care using electronic health services [10,30].

Table 1. Relationships between smart cities, digital public health, and One Digital Health (ODH).

Topics in common between (digital) public health and smart cities	ODH dimensions				
	Citizen engagement	Education	Human and veterinary health care	Industry 4.0	Surrounding environment
Surveillance					
Accidents and emergencies				✓	
Environmental conditions					✓
Electromagnetic radiation					✓
Health condition of older adults		✓	✓		
Emotions	✓				
Epidemics			✓		✓
Fitness activities	✓	✓	✓		
Food quality			✓	✓	✓
Healthy lifestyle promotion	✓	✓	✓		

Our main objective was to show and discuss how the ODH framework—combined with A&EI—can dynamically support the comprehensive monitoring and analysis of the daily life events of humans and animals in technologically integrated environments such as a smart home and a smart city.

The paper is structured as follows. After the introduction, we present and analyze a timely use case from a smart healthy city context [31]. We then describe the planning of an ODH intervention to deal with the use case features. Therefore, we remodel and update the original use case as an actual outcome of an ODH intervention. In the end, we introduce a part of the generic evaluation process of an ODH intervention and discuss its implementation potential. We conclude by pointing out the strengths and limitations of this use case, such as future developments of the ODH framework jointly with A&EI.

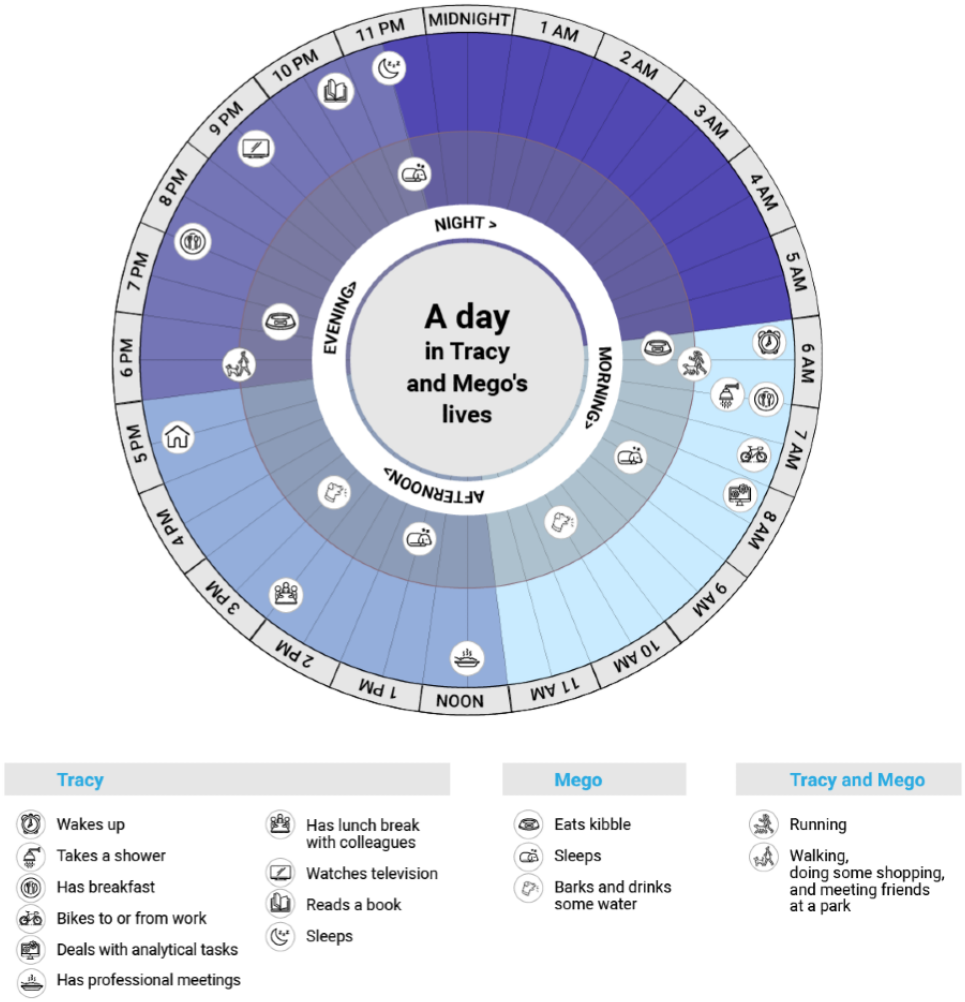
Use Case

Overview

A smart and healthy city well acknowledges several core ODH challenges. Our use case describes the sequence of steps needed to set up an ODH intervention around a user-centered demand and measure several parameters continuously, comprehensively, and reliably for effective multipurpose and integrative welfare monitoring. In this use case, we deal with human and animal health and welfare monitoring, continuously measuring physiological and behavioral parameters during their different activities—alone or together.

Tracy is a healthy single woman aged 27 years who shares her apartment in a metropolitan area with her dog Mego. Tracy is a junior finance analyst (ie, a tertiary position with responsibilities that might cause stress). During working weekdays, Tracy follows some typical activities of daily living. Mego is a healthy midsized male Shiba Inu dog aged 5 years. Mego lives with his owner Tracy. Most of Mego's time is spent at home (Figure 1).

Figure 1. A day in Tracy and Mego's lives.



A Regular Day in Tracy and Mego's Lives

At 6 AM, Tracy usually wakes up and checks that Mego's smart pet feeder delivered kibble and water. She then goes for a 30-minute run with Mego. Returning at 6:50 AM, Tracy takes a shower and has breakfast while briefly checking emails, social media, and the news. At 7:30 AM, she goes to the office by car or bicycle depending on the weather. Tracy works from 8 AM to 5:30 PM. She starts by planning and reviewing professional emails and then tackles high-priority tasks before handling analytical tasks. She takes a 1-hour lunch break with colleagues approximately at noon. The afternoon is organized around meetings and administrative tasks.

To return home, if Tracy used the bicycle on her way to work and there are issues on the way back, she may choose to take it on the bus or train. During this time, Mego, similar to other dogs, rests for a large part of the day and occasionally wakes up and barks to react to surrounding noises, plays with his toys, or drinks some water [13].

At around 6 PM, Mego warmly welcomes Tracy as he is excited to go out. If the weather, unexpected events, and mood conditions allow, they spend approximately 1 hour on a walk. During this time, Tracy does some shopping and meets friends in a park [32]. By 7:15 PM, they return home, and Mego receives his evening food ration. Tracy cooks dinner, watches television, reads a book, and prepares her plans for the next day. At 11:30 PM, Tracy goes to sleep.

Figure 1 shows a 24-slice clock in which the main activities of Tracy and Mego are reported depending on the moment of the day. In particular, an inner circle is recognizable that reports Mego's activities (eg, sleeping or eating kibble). An external circle features Tracy's activities (eg, moving to or from work or reading a book). Activities involving both of them are instead reported on the partition line between the aforementioned circles.

Some Important Spatial-Temporal Facts in Tracy and Mego's Daily Lives

Activity

Tracy spends most of her time indoors at home, her working place, shopping centers, restaurants, and the gym. Outdoor activities include running, working, hanging out with friends in parks and other open spaces, and using public transportation [33-35].

Tracy uses different means of transportation, such as walking, bicycle, car, and public vehicles. From home to work and back, she takes approximately 1 hour of biking—a considerable physical activity. Tracy's smartwatch informs her of caloric consumption, pulse, and oxygen saturation. The last physiological measurements are regularly stored in the cloud,

allowing Tracy to follow and share her physical activities [8]. Mego's sensor-endowed collar (an IoT device) connects to Tracy's smartwatch so that she can also monitor his activity and vitals. Most of the day, Mego rests (60% of the time; 14.5 hours a day). However, he has moderate activity (30% of the time; 7 hours a day), such as reacting to environmental stimuli (eg, welcoming Tracy when she returns home or barking at external noises). Mego has an intense physical activity (10% of the time; 2.5 hours a day) when he goes out for a walk or run with Tracy, plays with other dogs in the park, or is at home with his toys either alone or with Tracy [13].

Sleep and Rest

When going to bed, Tracy takes off her smartwatch for charging. This is also to prevent this device from disrupting her sleep quality. Tracy turns on her smart home speakers for relaxing music. She also bought a pair of blue light–blocking glasses to guarantee high-quality naps [36]. Mego rests most of the time [13] on a dog bed in Tracy's room.

Nutrition

Tracy has breakfast and dinner at home. Her smart fridge provides potential recipes with the available products [37,38]. She eats catering food at work, such as crudités, cooked legumes, and other protein sources. The cashier's receipt comprises all the related nutritional intakes [39-41].

Following the recommendation of the veterinary physician, Mego eats kibble. Tracy owns a sensor-endowed feeding station for effective feeding control and regulation [42]. In case of a product recall, Tracy receives a message on her mobile phone explaining the relevant "what to do's" [43].

A Particular Day in Tracy's Life—Example of a Health-Related Disruptive Event

One day, during a busier and more stressful period than regular at work, Tracy took her bicycle to work, but because of the weather, she took the train back home. Afterward, when walking in the park with Mego, Tracy suddenly collapsed. Unfortunately, she had not met her friends yet, nor had anyone happened to be around. Mego noticed the problem and started barking. After a while, some people came along, and finally, the ambulance arrived. Tracy stayed a few days in the hospital and another week at home to recover. During this time, Tracy's friends had been taking care of Mego, hosting him and trying to preserve his routine to reduce the negative impact of Tracy's absence (ie, during hospitalization) or unavailability to take care of him on her own (ie, during recovery).

Methods

Planning an ODH Intervention

An ODH intervention can be seen as the implementation of a set of digital functionalities, or digitalities [24], designed and deployed to (1) support specific initiatives that address human, animal, and environmental system needs and challenges; (2) assess and study these systems' expected and unexpected outcomes and effects and collect related data; and (3) select timely metrics for the outcomes of multicriteria decision analyses.

An ODH intervention is implemented to (1) *address One Health–related challenges*; (2) *achieve One Health–related* important and strategic *outcomes* for clinical follow-up and practice, such as for technology improvements needed; and (3) *achieve* FAIR uses of *digital technologies* [30,44].

In our case, this translates to (1) *a challenge* aiming to enhance aspects of healthy lifestyle promotion and surveillance; (2) *an outcome* consisting of performing effective monitoring of human and animal welfare within the context of a smart city, where health is a pivotal component; and (3) sensors for monitoring as *digital technologies* through which the intervention is implemented.

Effective monitoring of humans and animals in a smart environment must cover all relevant locations (indoors and outdoors) and define the critical parameters of health care and QoL [45]. Table 2 reports the 9 major types of interactions between the intervention recipients considering the extant strict interconnectedness between them [46]. Each cell reports instances of how one actor (row) positively or negatively affects another (column).

Table 2. Examples of interactions between the 3 One Health components.

	Human	Animal	Environment
Human	• Face to face [6] • Interactive (technology-based; eg, via social media) [8,47]	• Food, habitat provision, emotional bond, and health and well-being control [48] • Population follow-up and birth control [49,50] • Environment destruction [51]	• Compliance and sustainability for a technological environment at home, on a bicycle, in the steering wheel of a car, in the fridge handle, and more [52] • Urbanization [53-55] • Climate change, pollution, and regulation [51]
Animal	• Emotional and physical bond; health, well-being, and safety feelings (eg, dog) [56] • Disease vectors (zoonoses) [57]	• Food chain [58,59] • Natural regulation of populations [60]	• Soil fertilization [61,62] • Natural hazard (invasive species) [63-66]
Environment	• Food chain [58,59] • Natural hazards and disasters [67] • Space for well-being development [68-70]	• Availability of survival resources (food and shelter) [71] • Chemicals' influence on animal reproduction [72]	• Wildfire impacts on slope stability triggering in mountain areas [73] • Natural hazards and disasters [67]

Therefore, all the collected physiological, behavioral, and environmental data must (1) comprise indoor and outdoor locations where subjects are active and in contact [74,75], (2) be continuously time stamped, and (3) be shareable in a FAIR way [27].

Adverse Health Events

The ODH framework delivers a variety of observations for continuous health monitoring. It aims at analyzing health data on environmental, behavioral, physiological, and psychological domains, which is in line with the WHO QoL definition [76]. Continuous health monitoring allows continuous data analytics and even subtle trends to become recognizable at the early stages. This then opens many options for preventive medicine, hindering adverse health events (AHEs) [77]. However, AHEs will still occur but maybe not as frequently as without continuous monitoring. In such an event, the measurements taken from the subject are helpful. As of today, smart homes, smart cars, smart wearables, and smart clothes send out emergency calls, but so far, useful information—although available at the site—is not transferred because of lacking protocols and communication standards. This is addressed by A&EI [23]—in an emergency, whether an accident or a health-related adverse event, every second in fact counts as it carries data correlated with the individual involved. In a temporal-spatial continuous monitoring supported by the dynamic point of perception, the subject's associated data are distributed in each measuring smart environment and device. Upon the occurrence of an event, such isolated data silos have to concatenate to build the complete informative understanding of the latest health status of the subject. The International Standard Accident Number (ISAN) takes over the role, aiming at (1) standardizing an event by associating a unique number (token) composed of time and location of the event and ID number, (2) automatically collecting the corresponding isolated data slices of the individual from the alerting system (ie, smart environments and wearables), (3) automatically generating the alert and transferring it to the responding system (ie, emergency service), and (4) simultaneously transferring the vital and nonvital data to the curing system (ie, hospital) for informing the medical personnel before the subject is delivered to the hospital [78,79].

ODH Intervention Steps

Overview

We deconstruct the ODH intervention into single independent events and steps (globally called "activities") as items to be organized in tabular form, called the ODH intervention table (see Multimedia Appendix 1). For each activity, a number of fields must be identified, which relate to, for example, ODH domains and dimensions considered, digitalities involved, and their eventual mutual linkages (FAIRness levels).

Activity Identification

Each activity is assigned to a single row. The field "Activity UID" indicates the unique numerical identifier of the activity within the ODH intervention. The "Activity name" describes the single ODH intervention step. If the activity is broken down into subtasks, each one of them will be evaluated individually (ie, as a single row).

ODH Dimension Scores

An ODH dimension score assesses the dimensions of the ODH framework that directly relate to the activity. Each dimension obtains a specific importance value (increasing integer from 1 to 5) that indicates its connection rate to the corresponding activity of the intervention. The dimensions are reported as follows: C (*citizen engagement*), E (*education*), H (*human and veterinary health care*), I (*health care Industry 4.0* as health care services and technologies involving, eg, robots, 3D printing, cutting-edge ITs, and artificial intelligence [24,80]), and S (*surrounding environment*).

Main Digitality Domain

The field refers to 1 of the 3 areas of digital functionalities (humanities, animalities, and environmentalities) encompassed by the technology ring within the ODH steering wheel [24]. This field is divided into 2 subfields.

The first one, named "*Speciality*," reports the digitality explicitly deployed in the corresponding activity. Each speciality is assigned to a single row. This means that each single activity is operationalized by one or more specialities and is characterized by the marking letter of the domain it refers to: H (human), A (animal), or S (surrounding environment). Accordingly, H(Si) denotes the *i*th speciality that operationalizes the general digital functionality from the human domain, which delivers the corresponding activity.

The second subfield, named "*Technology*," refers instead to one or more technological solutions through which the speciality is deployed. Each technology is assigned a single row as well.

FAIR Data and Data FAIRness

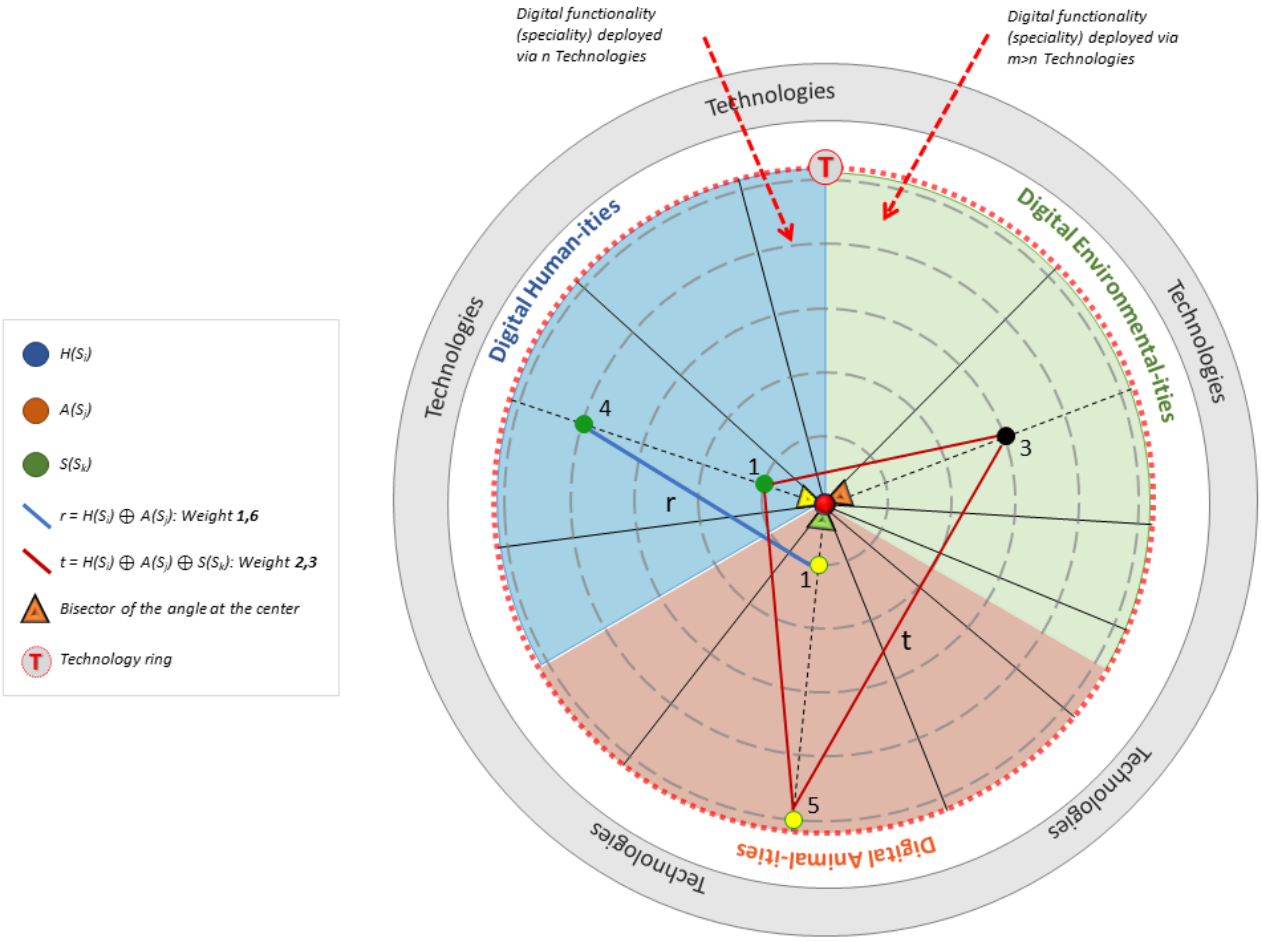
Assessing an ODH intervention's FAIRness is a basic requirement to correctly process the entire management cycle and stewardship of the data collected and shared during the intervention itself. Thus, the design and deployment steps of an ODH intervention need for its data to be FAIR. In this context, it means that the data must be (1) "Findable" by allowing for their discovery and sharing continuously between different monitoring, analyzing, and alerting systems; (2) "Accessible" by allowing the relevant and approved individuals and connected systems (ie, on behalf of accredited organizations) to deal with data and information when, where, and how needed to manage regular and disruptive events; (3) "Interoperable" by involving health-related communication, data exchanges, and processing standard protocols offered in a secured technological framework; and (4) "Reusable" to allow for a systematic, continuous, and intelligent integration of big multidimensional data for primary (eg, real-time clinical and environment monitoring) and secondary (eg, clinical and epidemiological follow-up) uses [29].

More specifically, a FAIR data assessment is conducted for every technology singled out so that an overall score is extracted for the speciality that the technologies refer to [29]. We mark satisfying, good, and total versus inadequate FAIRness of an individual technology with the symbols + or −, respectively.

Related Digitality Domain

This points out 1 of the other 2 areas of the technology ring (Figure 2) involved in the same activity as the main one. Similar to the main digitality domain, we divide related digitalities into the subfields *speciality* and *technology* and assess their level of FAIRness.

Figure 2. Example of a graphical representation of digitalities involved in a One Digital Health intervention.



Data Linkage

In total, 2 digitalities from 2 different domains form a relationship. If 3 digitalities—one from each of the 3 domains—connect with each other, we refer to it as a *triality*, and the field is duplicated. The **"**Data Linkage" field relates to the relationship or triality between the specialities described in the previous subsections. Equation 1 describes an example of a relationship between 2 specialities:

$$r_{HA} = H(S_i) \oplus A(S_j) \quad \textbf{(1)}$$

where r_{HA} is the relationship between specialities, $H(S_i)$ is the *i*th speciality related to a digital functionality in the human domain, $A(S_j)$ is the *j*th speciality related to a digital functionality in the animal domain, and ⊕ denotes a special defined operation similar to addition. It implies a possible confrontation between different specific deployments (in different domains) of the same kind of digital functionality.

In the absence of a relationship between one speciality from one domain and another speciality from another domain, the symbolism of equation 1 becomes as follows:

$$r_{H\emptyset} = H(S_i) \oplus \emptyset \quad \textbf{(2)}$$

where ø is the null element (there is no speciality connecting with S_i) and $r_{H\emptyset}$ is the absence of a relationship (the second member is null). Consequently, we represent a triality as:

$$t = r_{HAS} = H(S_i) \oplus A(S_j) \oplus S(S_k) \quad \textbf{(3)}$$

where $t = r_{HAS}$ is the triality (the specialities relate to all 3 domains) and $S(S_k)$ is the *k*th speciality related to a digital functionality in the environment domain.

Technology Ring

The technology ring serves as a catalyst among all the digital functionalities that the ODH framework relies on [24]. Accordingly, an ODH intervention takes place ideally within it through a series of steps described as follows. Each of the 3 domains is divided into as many circular sectors as digitalities are involved in the ODH intervention (Figure 2).

The width of each circular sector is directly proportional to the number of technologies deployed by the single speciality. Then, the bisector of the angle at the center is identified for each sector and represented by a dashed line. A dot is located along the bisector to represent the rate of involvement of the referring digitality within the ODH intervention.

The position of each dot is the output of the overall assessment of the FAIRness of the digitality considered, which involves all the possible technologies related at once; the lesser the output of the assessment, the farther the dot is from the center, and the higher the associated score (decreasing importance from 1 to 5; integer values) of the digitality.

We define the harmonic average of the associated mentioned scores as the weight of a relationship between 2 digitalities:

where M_g=M_g(S_i) is the harmonic average, N=2 is the number of digitalities related to each other, and S_i is the score associated with each of the specialities that operationalize the digital functionalities involved in the relationship.

We use a harmonic average as it is based on all observations, gives larger weights to smaller observations, and is thus more robust to strong observations and fluctuations of samples. Moreover, it can cope with variable time factors—in our case, the same speciality deployed via different technologies in different domains and in different time settings.

For a triality, we calculate the M_g value for each of the 3 single relationships and then their arithmetic average. M_g is a function of the scores associated with the 2 related dots. The higher the output of the FAIRness assessment of the digitalities, the closer the corresponding dot is to the center of the technology ring, the lower the ranking of the relationship (decreasing importance from 1 to 5), and the more effective that portion of the ODH intervention is. From a graphical point of view, the dots and relationships shall tend toward the center of the ring. For an optimal ODH intervention, its ODHness (ie, an overall evaluation of how well the ODH intervention is delivered) has to tend to 0 [24].

Eventually, the weight associated with the dimensions involved in the ODH intervention is a corrective factor that improves the scores obtained in the previous steps. Accordingly, the greater the number of dimensions involved in the ODH intervention, the more effective the corrective factor is supposed to be.

Ethical Considerations

Our research is not an observational study and does not produce or analyze any data taken from humans. Rather, we present a viewpoint and a use case as a potential scenario for an ODH intervention. Accordingly, a review by an ethics committee is not required.

However, planning an ODH intervention must cope with data privacy and security. In our opinion, confidentiality protection is a sensitive issue. In this use case, we do not focus specifically on this aspect. Nevertheless, in a real-world operationalization, it will be critical to have a dedicated data protection framework that allows us to deal when needed with data anonymization and deidentification.

In addition, anyone interested in running (primary use) or using data collected (secondary and tertiary uses) during an ODH intervention in real-world conditions must, according to the relevant local rules, obtain the approval of the relevant research ethics and data protection committees.

Results

Use Case (Updated): A Day in Tracy's Life in a Smart Healthy City

To picture this vision, we flashback to our use case with Tracy and Mego. This time, we tell the same story but set in a smart city using smart devices.

Tracy wakes up while her smart home processes the physiological measurements acquired overnight with IoT devices: a radar sensor on top of the bed monitors the respiratory rate, and the bed is equipped with a capacitive electrocardiograph in the mattress. Tracy also uses ballistocardiography and seismocardiography mounted on the bed frame for cardiorespiratory measurements and sleep assessment. When Tracy opens a door or presses a switch, her body temperature, electrodermal activity (EDA), and photoplethysmography (PPG) are taken, which deliver heart and respiratory rates upon touching the smart door handle or smart switch, respectively. In addition, passive infrared sensors gather activity data for both Tracy and Mego. In the bathroom, the smart toilet analyzes her urine and measures the pH value and density of excrements. In front of the mirror, Tracy's body temperature is measured by an infrared thermal sensor, whereas the weight scale integrated into records PPG and EDA. When watching television, Tracy's electrocardiogram (ECG) chair records her heart and respiratory rates.

Before Tracy leaves her apartment, she checks for Mego's general health status, provided by his smart dog collar, on her app. Upon Tracy's exit, her smart home transfers the health monitoring to her wearable sensors: Tracy's smart clothes are equipped with an embedded ECG. During all her outdoor activities, Tracy's wearable device performs the measurement, and data on the air quality are recorded [17,45].

Today, Tracy uses her car to drive to work. Opening the smart car door activates the measuring system, including body temperature, EDA, and ECG from the steering wheel as well as heart and respiratory rates from image-based PPG computed from the smart car's indoor camera. It is worth mentioning as well that Tracy's bicycle is equipped with a smart handlebar that also delivers health data.

A Particular Day in Tracy's Life (Updated)

Some days ago, the continuous health monitoring system reported to Tracy that she should consult a physician—the sequence and length of atrial fibrillation periods had slightly increased, a well-known harbinger of stroke. However, because of time conflicts, the physician's appointment was still outstanding when Tracy collapsed during her evening walk in the park with Mego. Instantaneously, Tracy's smartwatch detected the AHE and asked Tracy to release the alert return. While the countdown had not yet ended, Mego's collar reported Mego's extraordinary excitement to Tracy's smartwatch, which immediately generated the ISAN and sent an emergency alert. Using the ISAN, the rescue team requested access to the

smartwatch so as to check previous and ongoing measurements and, therefore, be well prepared when arriving at Tracy's location. This made it possible to perform the right intervention at the earliest point in time. Tracy recovered soon without staying in the hospital overnight.

Analyzing the ODH Intervention

Multimedia Appendix 1 showcases examples of "how-to" analysis of several "activities" (2 in regular time and 1 during a health-related disruptive event) as part of a wider ODH intervention. For each one of them, the rates of involvement of the ODH dimensions are scored, and the relevant technical fields are evaluated in light of the FAIR principles prism. Eventually, for each activity, the data linkage formula is also reported.

The first example relates to the monitoring of Tracy's respiratory rate via a radar sensor on top of her bed. From the A&EI viewpoint, this activity allows for the detection of respiratory abnormalities to be reported, if needed, in real time to an emergency medical service or be collected for future investigation.

The second example relates to the possibility for Tracy to check on the activity of the wearable devices she and Mego use. This may allow Tracy to know that, "in case of emergency" (eg, fever and muscular pain detected in Mego [before] and Tracy [after] may be signs of a zoonotic phenomenon, such as a bacterial infection known as leptospirosis), the system will be able to send relevant data from each of them to their own medical records [81].

In both cases, the industrial- and health care–related dimensions of the ODH framework appear to be highly involved. The surrounding environment is only slightly affected, whereas no notable involvement emerges for the human-related dimensions.

The last example involves all the ODH dimensions. In this case, using a smartwatch to send an alert message reflects some kind of citizen engagement (C) in sharing data with health care providers (M) that need to be trained (E) to use advanced technologies based on industrial standards (I) that can be used in different environments (S).

Discussion

Overview

In this viewpoint and use case, our main objective was to show and discuss how the ODH framework will support 24/7/365 technology-based health and environment monitoring to provide the right answer to the right event (related to a human, animal, or place) in the right place at the right time and with the right means. The central challenge in applying ODH in real-world conditions is the integration of digital development and technologies into 1 health concept. This comes with their use to achieve One Health goals rather than redefining and reconceptualizing One Health in the face of technological advancements [10,11,24,29]. In this regard, planning an ODH intervention is critical for taking forward the integration of the 3 main One Health domains (human, animal, and ecosystem) in light of their digital and computational components. The analyzed use case emerges from the combination of smart cities and digital public health. It suggests an around-the-clock scenario of sensor-based welfare monitoring for a human and a pet in a smart environment context—a large part of Tracy and Mego's lives involves technological measuring and monitoring systems related to IoMT or IoAHT [8,12]. To enhance the understanding of the impact of an ODH intervention and the subsequent assessment of its ODHness, we proposed a way to analyze human and animal activities in different environments by quantifying their relationships with the ODH framework features [24] and their links to technical fields by considering the assessment of the FAIR guiding principles for the 3 areas identified and included within the ODH technology ring [29,76].

Principal Findings

In the use case supporting our viewpoint, as in real life, the large amount of data generated by IoT sensors and technologies allows for an effective analysis of behaviors, habits, pattern extraction, and medical conditions of different types of subjects in the mid- and long term. In addition to the initial aim of health condition prediction, prevention, and early alert perspective—usually reported in the existing literature for single contexts such as home [82], hospital [83], or even veterinary epidemiology [84] and that ODH already addresses in a unified way—a growing importance has arisen for disruptive events and punctual abnormalities.

For Tracy, this could be related, for instance, to a medical emergency, such as dyspnea or cardiac arrest during sleep or a physical activity, detected by physiological sensors. It can be, as another example, a bicycle accident detected by physiological sensors, accelerometers, and automatic alerts to emergency services. An additional example relates to contact tracing via an app for highly contagious diseases such as COVID-19, which can alert Tracy and support her in making timely decisions according to the public health authorities' recommendations.

For Mego, a disruptive event can be a veterinary emergency in case of an accident, as seen for Tracy, or a change in the hydration frequency that can suggest a food-based intoxication.

Moreover, a disruptive and dangerous event can involve both Tracy and Mego, such as a fire, with Tracy's physiological parameters dramatically changing within a couple of minutes overnight and Mego's similarly agitated behavior.

Such aspects of emergency management demand a continuous follow-up of humans', animals', and surrounding environments' health care in a P5 medicine approach [85-88] that combines (1) prediction, acquiring data and building relevant models supporting emergency and disaster preparedness [8]; (2) prevention, detecting (weak) signals of abnormalities and treating them before strong deviations (eg, preventive confinement to avoid an epidemic); (3) participation, involving the handling of human, animal, or environmental issues (eg, engaging in vaccination campaigns [89-91]); (4) personalization, proposing the use of the adapted solutions to a detected issue (a drug-based treatment for a human or an animal or the development of a repopulation program for a hurt vegetable ecosystem); and (5) precision, delivering the right intervention at the right time (as much as possible in real time) by the right

individual on the right ecosystem components to get to, for example, efficient and dedicated evacuation plans [92,93].

In addition, a specific interpretation is also delivered for what concerns (1) citizen engagement and education aimed at information development and large-scale information gathering (so-called citizen science [94]) for first aid, basic health-related technologies, and sustainability using, for example, mobile apps to increase their penetration rate in the grand public and (2) interoperability from an industrial perspective to facilitate the interconnection and communication between the different IoT systems and alert systems to enhance end-to-end accident and disaster management [78,79].

As shown, singling out the activities that the ODH intervention in a "smart city meets public health" scenario would be made up of and working the technology ring out accordingly (also) to define its ODHness aimed to commence the dialogue with topics from different perspectives dealing with health communication (eg, health promotion based on engagement and education [6,95]) and surveillance (eg, monitoring health conditions and physical activity, detecting and preparing the health care system for disruptive and long-term events, and behavior or environmental changes) [9].

A&EI discipline showed its contributions to the use case regarding the short-term and abrupt event and abnormality detection as well as long-term prediction and prevention. The use-case analysis indicated that A&EI aims to (1) turn smart private spaces into diagnostic spaces unobtrusively (eg, home and car) [96], (2) build continuous measurement and monitoring via dynamic points of perception, and (3) achieve interconnectivity and communication of the means of measurement. The field focuses on the data linkage and interconnectivity of medical and nonmedical sensors and devices in smart environments to construct and support the onboard and distributed data processing and analysis in hierarchical levels of abstracts [12]. The integration of ODH and A&EI in this use case contributes to the development of educational projects and programs, allowing health, environment, engineering, design, and business students and trainees to develop their creative and critical abilities by proposing new concepts and systems [42,95,97]. These education projects could be used in future smart cities wherein One Health and disaster (ie, accidents and emergencies) prevention and preparedness will be daily life pillars [98,99]. This aspect fits with the United Nations 2030 Agenda for Sustainable Development, which points out that young people should be educated in such a way that smart health monitoring is not only for unhealthy humans and older adults [100,101]. For example, analyzing walking steps is a default functionality of mobile devices and is used by young and healthy people to measure their physical activity in real time. Moreover, from an A&EI perspective, preventing complications of a potential medical emergency is something that smartwatches have yet to achieve by detecting and alerting their owners of the first signs of a cardiac event [102].

What emerges is that the very act of planning an ODH intervention placed particular emphasis on aspects of emergency management, which in this context refers directly back to A&EI. This is actually a way to deal with the same dimensions as ODH. Therefore, it is plausible to say that the ODH framework can be enriched with a new layer-like cross-sectional element so that an infinitely recurring loop is created between the 2 models.

In summary, ODH and A&EI as a whole contribute to enhancing (1) the overall QoL of the smart city inhabitants (both humans and animals); (2) the public (digital) health policies and processes frame, such as the entire smart city ecosystem development and management, that is, architecture (eg, accessibility for people with disabilities and reduction of energy consumption) and urbanization (eg, communication systems, transportation networks, malls, education, and health care center locations); and (3) the communication between the different health care system actors involved, such as clinicians, engineers, regulators, and administrators and, more generally, the 3 domains that the ODH framework is made of. The existing Medical Informatics and Digital Health Multilingual Ontology could be expanded and adapted to this end [103,104].

Viewpoint Constraints

A correct development of the "smart city meets digital public health" field is currently still hindered by a number of factors, as briefly discussed in the following sections.

Governance

The control and governance of environmental sustainability can be best approached by assessing ecosystem services that are capable of quantifying and valuing all the goods and services that are generated within the ecosystems themselves. Recently, such networks have been increasingly endowed with digital technologies such as (1) environmental management and monitoring information systems; (2) automated and scalable approaches for collecting, digitalizing, and assembling geocoded big data; and (3) information-fusion algorithms and artificial intelligence that use multiple data streams and clinical decision support algorithms that integrate population-based, public health–focused perspectives into outbreak detection–focused management systems [45,105,106].

Data Security and Privacy

However, it is necessary to consider that using electronic and computational systems to collect, store, and analyze data and react and deliver an appropriate response induces at least 2 challenges. Dealing with health care customer data requires dealing with their security and privacy from an ethical viewpoint to ensure patient rights [107] and prevent data theft that can induce, for example, health-related device hijacking with dramatic consequences [108]. Moreover, the health-related data acquisition and monitoring systems used in smart cities demand seamless and secured bidirectional communication, supporting high-speed, large-bandwidth, and low-latency internet infrastructures materialized by the fifth and sixth generation of wireless network technology (respectively known as 5G and 6G) and future versions. Moreover, the use of cloud-based storage and computational resources is essential to allow (near) real-time data collection and analysis.

Sustainability

Accordingly, from a sustainability perspective, as smart cities are electric power–based and digital green, digital health green

infrastructures must be developed to reduce the environmental fingerprint [109,110].

Therefore, from an industrial perspective, reducing power consumption, electronic waste, and biodegradable health care–related materials is a global challenge [111,112].

Interoperability

Furthermore, in smart cities, the heterogeneity of IoT devices, and specifically those related to the health care industry, must benefit from improved interoperability standards and procedures and device compatibility. This last point must critically affect the health service providers' and customers' acceptance and use of mobile and ubiquitous technologies [113].

Health Communication and Disparity

Nevertheless, a limitation of having a generalized use of IoT devices or, more globally, a healthy smart city lies on health communication and disparity. Indeed, using IoT may be challenging for different groups such as older adults as well as individuals with communication disabilities or limited abilities. Accordingly, even though IoT is taking health care to the next stage worldwide, it could paradoxically exclude some people who need it [6,114]. Therefore, a similar threat may affect the veterinary sector, whose development rate is still quite lagging behind that of human medicine.

Ethical and Cultural Limitations

In addition, collecting continuous data on anyone and anything may entail ethical and cultural limitations both related to privacy. In a smart city, data collection and processing are performed such that the customers of health care services are not (fully) aware of the sharing of personal data [115]. Moreover, developing a smart city wherein health monitoring is a pillar requires a high technology acceptance level when we consider that technologies are everywhere. Just to mention another critical sector such as transportation safety, the development of IoT-based vehicle accident detection is important, but privacy limitations exist regarding the fact that any driving event can be recorded even if it is not related to an accident [116,117].

In the use case presented, the use of the different activity trackers points out that the technology is accepted by our persona (Tracy), yet it is an open question whether she was fully aware of "giving up" part of her privacy rights [4,118].

Limitations of the Study

Some limitations of this work arise from the way the use case for the ODH intervention was determined. First, the "Animal" domain was only represented by a pet. The presence or role of other kinds of animals—such as nonconventional pets (eg, rabbits and reptiles), herds (eg, cows and goats), or wildlife (eg, wolves and boars)—was not addressed. In addition, the "Environment" domain was not properly included in the use case, whereas in a smart and healthy city, the management of environmental resources, made itself smart via IoT and information and communications technologies, is a critical element toward actual e-governance policy planning [119].

Furthermore, a major focus was on the general category of IoT that, although an all-encompassing type of technology, is currently dealing with peculiar declinations for what concerns the general (human) health care sector with IoMT and, in more recent times, animals' health and well-being status with IoAHT [120].

Another findable limitation is that the proposed use case only describes a single day in Tracy and Mego's lives. Although Tracy's working days are characterized by many routine activities, the analysis of a longer period would have likely involved many more things to be discussed.

Future Perspectives

Tracy's daily occupations and activities presented in this fictional use case are applicable to a large part of the population living in similar contexts.

One of the potential and expected implications of ODH interventions lies on supporting digital health literacy, from children to older adults, to engage with personal health, public health, and environmental monitoring systems to increase awareness. Improvement in this aspect is likely to lead to better health outcomes and a more proactive approach to medical practice, thus reducing both the digital divide and health inequalities [6,121].

On the basis of these assumptions, the development of policies regarding health data management and sharing for secondary use [122-126] is expected to be facilitated and perhaps automated. A timely and FAIR deployment of ODH- and A&EI-based data collection modalities would lead to the establishment of more comprehensive guidelines and decision support systems for what concerns the many (interrelated) features of health care and to a better understanding of user expectations.

The combination of health literacy, human expectations, and data collection—or, rather, the logical flow through these points—yields scenarios where information is stored in electronic health records in a timely manner and available in health care management organizations' repositories and allows for the implementation of personalized medicine processes considering environmental measurements and human [127] or animal [120,128,129] behaviors. In a broader, holistic view, it looks at a proactive medical practice that no longer addresses patients but humans outside the human-centered health care systems whose improved awareness even steps beyond the concept of "empowerment" [130,131]. This means that electronic health record data shall be processed with other data such as lifestyle habit data, for example, patient-generated health data (ie, not reported to health care professionals) and environmental data (eg, pet owning [42] and hobbies). An integrative understanding of the human way of life and human needs and expectations will help efficiently and effectively enhance health communication–targeted campaigns to improve disease prevention, detection, and follow-up at a large scale [6].

Conclusions

Our vision for the future sees smart cities as an emerging paradigm involving a large set of technologies (eg, IoT and, thus, digital health, telehealth, mobile health, and web-based social networks of patients and caregivers) and behavior-changing tools to refine education, engagement, the consumption of food, physical activity, and the use of technology. The main aim is to improve QoL and life expectancy. This has shifted health care focus from treatment after diagnosis to prediction and prevention and from a health care professional–centered (so-called paternalistic [132]) approach to a health care service user–centered follow-up management. Thus, health care is no longer limited to the walls of the health care centers (eg, clinics and hospitals) and the "homespitals" (also known as hospital at home) but is ubiquitous and based on continuous, comprehensive, and reliable measurement of several physiological and behavioral parameters. The integration of the ODH and A&EI viewpoints will allow for the reduction of disparities and loss of time in managing disruptive health-related events and for looking at health as a whole, wherein human and animal well-being in a secure and proactive environment. ODH and A&EI are triggers for developing and implementing precision public health, currently defined as imaginary, by dealing with the entire data management process from end to end [11].

Acknowledgments

The authors are grateful to Mrs Orly Seligmann (Bachelor of Design), visual communication designer at the Holon Institute of Technology, for her assistance.

Data Availability

Data sharing is not applicable to this paper as no data sets were generated or analyzed during this study.

Authors' Contributions

AB contributed to conceptualization, formal analysis, funding acquisition, investigation, methodology, project administration, supervision, validation, visualization, and writing (original draft, review, and editing). MH contributed to conceptualization, formal analysis, investigation, methodology, validation, and writing (original draft, review, and editing). TMD contributed to methodology, validation, and writing (review and editing). OT contributed to conceptualization, formal analysis, investigation, methodology, supervision, validation, and writing (original draft, review, and editing).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Excerpts of the One Digital Health (ODH) intervention table for this use case.
[PDF File (Adobe PDF File), 252 KB - medinform_v11i1e43871_app1.pdf]

References

1. Pentland A. The data-driven society. Sci Am 2013 Oct;309(4):78-83. [doi: 10.1038/scientificamerican1013-78] [Medline: 24137860]
2. Angelidou M. Smart city policies: a spatial approach. Cities 2014 Jul;41(Supplement 1):S3-11 [FREE Full text] [doi: 10.1016/j.cities.2014.06.007]
3. Su K, Li J, Fu H. Smart city and the applications. In: Proceedings of the 2011 International Conference on Electronics, Communications and Control. 2011 Presented at: ICECC '11; September 9-11, 2011; Ningbo, China p. 1028-1031. [doi: 10.1109/icecc.2011.6066743]
4. Luzi D, Pecoraro F, Tamburis O. Appraising healthcare delivery provision: a framework to model business processes. Stud Health Technol Inform 2017;235:511-515. [Medline: 28423845]
5. Aversano L, Bernardi M, Cimitile M, Pecori R. Early detection of Parkinson disease using deep neural networks on gait dynamics. In: Proceedings of the 2020 International Joint Conference on Neural Networks. 2020 Presented at: IJCNN '20; July 19-24, 2020; Glasgow, UK p. 1-8 URL: https://ieeexplore.ieee.org/document/9207380 [doi: 10.1109/ijcnn48605.2020.9207380]
6. Benis A, Barak Barkan R, Sela T, Harel N. Communication behavior changes between patients with diabetes and healthcare providers over 9 years: retrospective cohort study. J Med Internet Res 2020 Aug 11;22(8):e17186 [FREE Full text] [doi: 10.2196/17186] [Medline: 32648555]
7. Ghazal TM, Hasan MK, Alshurideh MT, Alzoubi HM, Ahmad M, Akbar SS, et al. IoT for smart cities: machine learning approaches in smart healthcare—a review. Future Internet 2021 Aug 23;13(8):218 [FREE Full text] [doi: 10.3390/fi13080218]
8. Benis A. Social media and the internet of things for emergency and disaster medicine management. Stud Health Technol Inform 2022 May 20;291:105-117. [doi: 10.3233/SHTI220011] [Medline: 35593760]

9. Rocha NP, Dias A, Santinha G, Rodrigues M, Queirós A, Rodrigues C. Smart cities and public health: a systematic review. Procedia Comput Sci 2019;164:516-523 [FREE Full text] [doi: 10.1016/j.procs.2019.12.214]
10. Soobiah C, Cooper M, Kishimoto V, Bhatia RS, Scott T, Maloney S, et al. Identifying optimal frameworks to implement or evaluate digital health interventions: a scoping review protocol. BMJ Open 2020 Aug 13;10(8):e037643 [FREE Full text] [doi: 10.1136/bmjopen-2020-037643] [Medline: 32792444]
11. Wienert J, Jahnel T, Maaß L. What are digital public health interventions? First steps toward a definition and an intervention classification framework. J Med Internet Res 2022 Jun 28;24(6):e31921. [doi: 10.2196/31921] [Medline: 35763320]
12. Haghi M, Benis A, Deserno TM. Accident and emergency informatics and one digital health. Yearb Med Inform 2022 Aug;31(1):40-46 [FREE Full text] [doi: 10.1055/s-0042-1742506] [Medline: 35654425]
13. Griss S, Riemer S, Warembourg C, Sousa FM, Wera E, Berger-Gonzalez M, et al. If they could choose: how would dogs spend their days? Activity patterns in four populations of domestic dogs. Appl Anim Behav Sci 2021 Oct;243:105449 [FREE Full text] [doi: 10.1016/j.applanim.2021.105449]
14. Rock MJ, Adams CL, Degeling C, Massolo A, McCormack GR. Policies on pets for healthy cities: a conceptual framework. Health Promot Int 2015 Dec;30(4):976-986 [FREE Full text] [doi: 10.1093/heapro/dau017] [Medline: 24694682]
15. Di Martino B, Li KC, Yang LT, Esposito A. Internet of Everything: Algorithms, Methodologies, Technologies and Perspectives. Singapore, Singapore: Springer; 2018.
16. Tamburis O, Mangia M, Contenti M, Mercurio G, Rossi Mori A. The LITIS conceptual framework: measuring eHealth readiness and adoption dynamics across the healthcare organizations. Health Technol 2012 Apr 13;2(2):97-112 [FREE Full text] [doi: 10.1007/s12553-012-0024-5]
17. Lepenies R, Zakari IS. Citizen science for transformative air quality policy in Germany and Niger. Sustain 2021 Apr 02;13(7):3973 [FREE Full text] [doi: 10.3390/su13073973]
18. Ramaswami A, Russell AG, Culligan PJ, Sharma KR, Kumar E. Meta-principles for developing smart, sustainable, and healthy cities. Science 2016 May 20;352(6288):940-943. [doi: 10.1126/science.aaf7160] [Medline: 27199418]
19. Mouton M, Ducey A, Green J, Hardcastle L, Hoffman S, Leslie M, et al. Towards 'smart cities' as 'healthy cities': health equity in a digital age. Can J Public Health 2019 Jun;110(3):331-334 [FREE Full text] [doi: 10.17269/s41997-019-00177-5] [Medline: 30701413]
20. McConnell AR, Brown CM, Shoda TM, Stayton LE, Martin CE. Friends with benefits: on the positive consequences of pet ownership. J Pers Soc Psychol 2011 Dec;101(6):1239-1252. [doi: 10.1037/a0024506] [Medline: 21728449]
21. Malkina-Pykh IG, Pykh YA. Quality-of-life indicators at different scales: theoretical background. Ecol Indic 2008 Nov;8(6):854-862 [FREE Full text] [doi: 10.1016/j.ecolind.2007.01.008]
22. Hacker ED. Technology and quality of life outcomes. Semin Oncol Nurs 2010 Feb;26(1):47-58 [FREE Full text] [doi: 10.1016/j.soncn.2009.11.007] [Medline: 20152578]
23. Deserno TM, Haghi M, Al-Shorbaji N. Accident and Emergency Informatics. Washington, DC, USA: IOS Press; 2022.
24. Benis A, Tamburis O, Chronaki C, Moen A. One digital health: a unified framework for future health ecosystems. J Med Internet Res 2021 Feb 05;23(2):e22189 [FREE Full text] [doi: 10.2196/22189] [Medline: 33492240]
25. Hardy E, Standley CJ. Identifying intersectional feminist principles in the One Health framework. One Health 2022 Dec;15:100404 [FREE Full text] [doi: 10.1016/j.onehlt.2022.100404] [Medline: 35677572]
26. Fatehi F, Samadbeik M, Kazemi A. What is digital health? Review of definitions. Stud Health Technol Inform 2020 Nov 23;275:67-71. [doi: 10.3233/SHTI200696] [Medline: 33227742]
27. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR guiding principles for scientific data management and stewardship. Sci Data 2016 Mar 15;3:160018 [FREE Full text] [doi: 10.1038/sdata.2016.18] [Medline: 26978244]
28. Wilkinson MD, Sansone SA, Schultes E, Doorn P, Bonino da Silva Santos LO, Dumontier M. A design framework and exemplar metrics for FAIRness. Sci Data 2018 Jun 26;5:180118 [FREE Full text] [doi: 10.1038/sdata.2018.118] [Medline: 29944145]
29. Tamburis O, Benis A. One digital health for more FAIRness. Methods Inf Med 2022 Dec;61(S 02):e116-e124 [FREE Full text] [doi: 10.1055/a-1938-0533] [Medline: 36070786]
30. Classification of digital health interventions v1.0: a shared language to describe the uses of digital technology for health. World Health Organization. 2018 Mar. URL: https://apps.who.int/iris/bitstream/handle/10665/260480/WHO-RHR-18.06-eng.pdf?sequence=1&isAllowed=y [accessed 2022-10-28]
31. Solanas A, Patsakis C, Conti M, Vlachos IS, Ramos V, Falcone F, et al. Smart health: a context-aware health paradigm within smart cities. IEEE Commun Mag 2014 Aug;52(8):74-81 [FREE Full text] [doi: 10.1109/MCOM.2014.6871673]
32. Sanfridsson J, Sparrevik J, Hollenberg J, Nordberg P, Djärv T, Ringh M, et al. Drone delivery of an automated external defibrillator - a mixed method simulation study of bystander experience. Scand J Trauma Resusc Emerg Med 2019 Apr 08;27(1):40 [FREE Full text] [doi: 10.1186/s13049-019-0622-6] [Medline: 30961651]
33. Leech JA, Nelson WC, Burnett RT, Aaron S, Raizenne ME. It's about time: a comparison of Canadian and American time-activity patterns. J Expo Anal Environ Epidemiol 2002 Nov;12(6):427-432. [doi: 10.1038/sj.jea.7500244] [Medline: 12415491]

34. Brasche S, Bischof W. Daily time spent indoors in German homes--baseline data for the assessment of indoor exposure of German occupants. Int J Hyg Environ Health 2005;208(4):247-253. [doi: 10.1016/j.ijheh.2005.03.003] [Medline: 16078638]
35. Anthes E. The Great Indoors: The Surprising Science of How Buildings Shape Our Behavior, Health, and Happiness. New York, NY, USA: Farrar, Straus and Giroux; 2020.
36. Perez-Pozuelo I, Zhai B, Palotti J, Mall R, Aupetit M, Garcia-Gomez JM, et al. The future of sleep health: a data-driven revolution in sleep science and medicine. NPJ Digit Med 2020 Mar 23;3:42 [FREE Full text] [doi: 10.1038/s41746-020-0244-4] [Medline: 32219183]
37. Luo S, Jin JS, Li J. A smart fridge with an ability to enhance health and enable better nutrition. Int J Multimed Ubiquitous Eng 2009;4(2):69-80 [FREE Full text]
38. Smetana S, Aganovic K, Heinz V. Food supply chains as cyber-physical systems: a path for more sustainable personalized nutrition. Food Eng Rev 2020 Aug 22;13(1):92-103 [FREE Full text] [doi: 10.1007/s12393-020-09243-y]
39. Silvergreens restaurant is first to offer nutritional information printed on the back of a two-sided receipt. National Cash Register Corporation. 2018 Dec 18. URL: https://investor.ncr.com/news-releases/news-release-details/silvergreens-restaurant-first-offer-nutritional-information [accessed 2022-10-22]
40. Dumanovsky T, Huang CY, Nonas CA, Matte TD, Bassett MT, Silver LD. Changes in energy content of lunchtime purchases from fast food restaurants after introduction of calorie labelling: cross sectional customer surveys. BMJ 2011 Jul 26;343:d4464 [FREE Full text] [doi: 10.1136/bmj.d4464] [Medline: 21791497]
41. Should your supermarket receipt count calories? The Guardian. 2017. URL: https://www.theguardian.com/business/shortcuts/2017/jul/10/should-your-supermarket-receipt-count-calories [accessed 2022-10-28]
42. Benis A. Healthcare informatics project-based learning: an example of a technology management graduation project focusing on veterinary medicine. Stud Health Technol Inform 2018;255:267-271. [doi: 10.3233/978-1-61499-921-8-267] [Medline: 30306950]
43. Olsen P, Borit M. The components of a food traceability system. Trends Food Sci Technol 2018 Jul;77:143-149 [FREE Full text] [doi: 10.1016/j.tifs.2018.05.004]
44. Richardson S, Lawrence K, Schoenthaler AM, Mann D. A framework for digital health equity. NPJ Digit Med 2022 Aug 18;5(1):119 [FREE Full text] [doi: 10.1038/s41746-022-00663-0] [Medline: 35982146]
45. Tramontano A, Scala M, Magliulo M. Wearable devices for health-related quality of life evaluation. Soft Comput 2019 Jul 9;23(19):9315-9326 [FREE Full text] [doi: 10.1007/s00500-019-04123-y]
46. Hochmuth A, Exner AK, Dockweiler C. Implementierung und partizipative Gestaltung digitaler Gesundheitsinterventionen. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2020 Feb;63(2):145-152. [doi: 10.1007/s00103-019-03079-6] [Medline: 31938837]
47. Hu H, Galea S, Rosella L, Henry D. Big data and population health: focusing on the health impacts of the social, physical, and economic environment. Epidemiology 2017 Nov;28(6):759-762. [doi: 10.1097/EDE.0000000000000711] [Medline: 28682850]
48. Mellor DJ, Beausoleil NJ, Littlewood KE, McLean AN, McGreevy PD, Jones B, et al. The 2020 Five Domains Model: including human-animal interactions in assessments of animal welfare. Animals (Basel) 2020 Oct 14;10(10):1870 [FREE Full text] [doi: 10.3390/ani10101870] [Medline: 33066335]
49. Rutberg AT. Managing wildlife with contraception: why is it taking so long? J Zoo Wildl Med 2013 Dec;44(4 Suppl):S38-S46. [doi: 10.1638/1042-7260-44.4S.S38] [Medline: 24437084]
50. Garcia RC, Amaku M, Biondo AW, Ferreira F. Dog and cat population dynamics in an urban area: evaluation of a birth control strategy. Pesq Vet Bras 2018 Mar;38(3):511-518 [FREE Full text] [doi: 10.1590/1678-5150-pvb-4205]
51. Fitzmaurice M. Environmental compliance control. Wroclaw Rev Law Adm Econ 2018;8(2):372-394 [FREE Full text] [doi: 10.1515/wrlae-2018-0054]
52. Rogers DT. Environmental Compliance and Sustainability: Global Challenges and Perspectives. Boca Raton, FL, USA: CRC Press; 2020.
53. Phillips DR. Urbanization and human health. Parasitology 1993;106 Suppl:S93-107. [doi: 10.1017/s0031182000086145] [Medline: 8488075]
54. McKinney ML. Urbanization, biodiversity, and conservation: the impacts of urbanization on native species are poorly studied, but educating a highly urbanized human population about these impacts can greatly improve species conservation in all ecosystems. BioScience 2002 Oct;52(10):883-890 [FREE Full text] [doi: 10.1641/0006-3568(2002)052[0883:ubac]2.0.co;2]
55. Ahmed Z, Zafar MW, Ali S, Danish. Linking urbanization, human capital, and the ecological footprint in G7 countries: an empirical analysis. Sustain Cities Soc 2020 Apr;55:102064 [FREE Full text] [doi: 10.1016/j.scs.2020.102064]
56. Wells DL. The effects of animals on human health and well-being. J Soc Issues 2009 Jul 23;65(3):523-543 [FREE Full text] [doi: 10.1111/j.1540-4560.2009.01612.x]
57. Gortázar C, de la Fuente J. COVID-19 is likely to impact animal health. Prev Vet Med 2020 Jul;180:105030 [FREE Full text] [doi: 10.1016/j.prevetmed.2020.105030] [Medline: 32447153]
58. Briand F, Cohen JE. Environmental correlates of food chain length. Science 1987 Nov 13;238(4829):956-960. [doi: 10.1126/science.3672136] [Medline: 3672136]

59. Ward CL, McCann KS. A mechanistic theory for aquatic food chain length. Nat Commun 2017 Dec 11;8(1):2028 [FREE Full text] [doi: 10.1038/s41467-017-02157-0] [Medline: 29229910]
60. Weiler N. Do big carnivores practice birth control? Science. 2015 Apr 09. URL: https://www.science.org/content/article/do-big-carnivores-practice-birth-control [accessed 2022-10-28]
61. Guo Z, Zhang J, Fan J, Yang X, Yi Y, Han X, et al. Does animal manure application improve soil aggregation? Insights from nine long-term fertilization experiments. Sci Total Environ 2019 Apr 10;660:1029-1037. [doi: 10.1016/j.scitotenv.2019.01.051] [Medline: 30743900]
62. Ferreira PA, Ceretta CA, Lourenzi CR, De Conti L, Marchezan C, Girotto E, et al. Long-term effects of animal manures on nutrient recovery and soil quality in acid typic hapludalf under no-till conditions. Agron 2022 Jan 19;12(2):243 [FREE Full text] [doi: 10.3390/agronomy12020243]
63. Mooney HA, Cleland EE. The evolutionary impact of invasive species. Proc Natl Acad Sci U S A 2001 May 08;98(10):5446-5451 [FREE Full text] [doi: 10.1073/pnas.091093398] [Medline: 11344292]
64. Lemke A, Kowarik I, von der Lippe M. How traffic facilitates population expansion of invasive species along roads: the case of common ragweed in Germany. J Appl Ecol 2018 Nov 07;56(2):413-422 [FREE Full text] [doi: 10.1111/1365-2664.13287]
65. Brzeziński M, Żmihorski M, Zarzycka A, Zalewski A. Expansion and population dynamics of a non-native invasive species: the 40-year history of American mink colonisation of Poland. Biol Invasions 2018 Sep 12;21(2):531-545 [FREE Full text] [doi: 10.1007/s10530-018-1844-7]
66. Fraser EJ, Lambin X, Travis JM, Harrington LA, Palmer SC, Bocedi G, et al. Range expansion of an invasive species through a heterogeneous landscape - the case of American mink in Scotland. Diversity Distrib 2015 Jan 28;21(8):888-900 [FREE Full text] [doi: 10.1111/ddi.12303]
67. Hyndman D, Hyndman D. Natural Hazards and Disasters. 5th edition. Boston, MA, USA: Cengage Learning; 2016.
68. Hiramatsu Y, Ito A, Luo J, Hasegawa M, Sasaki A. Developing an application for walking in nature for post COVID-19. In: Proceedings of the 4th International Conference on Intelligent Human Systems Integration. 2021 Presented at: IHSI '21; February 22-24, 2021; Palermo, Italy p. 721-727. [doi: 10.1007/978-3-030-68017-6_107]
69. Choi S, Kim I. Sustainability of nature walking trails: predicting walking tourists' engagement in pro-environmental behaviors. Asia Pac J Tour Res 2021;26(7):748-767 [FREE Full text] [doi: 10.1080/10941665.2021.1908385]
70. Kotera Y, Lyons M, Vione KC, Norton B. Effect of nature walks on depression and anxiety: a systematic review. Sustain 2021 Apr 04;13(7):4015 [FREE Full text] [doi: 10.3390/su13074015]
71. Thies W, Kalko EK, Schnitzler HU. Influence of environment and resource availability on activity patterns of Carollia Castanea (Phyllostomidae) in Panama. J Mammal 2006 Apr;87(2):331-338 [FREE Full text] [doi: 10.1644/05-mamm-a-161r1.1]
72. Sehonova P, Svobodova Z, Dolezelova P, Vosmerova P, Faggio C. Effects of waterborne antidepressants on non-target animals living in the aquatic environment: a review. Sci Total Environ 2018 Aug 01;631-632:789-794. [doi: 10.1016/j.scitotenv.2018.03.076] [Medline: 29727988]
73. Abbate A, Longoni L, Ivanov VI, Papini M. Wildfire impacts on slope stability triggering in mountain areas. Geosciences 2019 Sep 25;9(10):417 [FREE Full text] [doi: 10.3390/geosciences9100417]
74. Herrera F, Oh SY, Bailenson JN. Effect of behavioral realism on social interactions inside collaborative virtual environments. Presence (Camb) 2018 Feb 1;27(2):163-182 [FREE Full text] [doi: 10.1162/pres_a_00324]
75. Soltani S, Gu N, Ochoa JJ, Sivam A. The role of spatial configuration in moderating the relationship between social sustainability and urban density. Cities 2022 Feb;121:103519 [FREE Full text] [doi: 10.1016/j.cities.2021.103519]
76. The World Health Organization Quality of Life (WHOQOL). World Health Organization. 2012 Mar 01. URL: https://www.who.int/publications/i/item/WHO-HIS-HSI-Rev.2012.03 [accessed 2022-10-28]
77. Cafri G, Li L, Paxton EW, Fan J. Predicting risk for adverse health events using random forest. J Appl Stat 2018 Sep 10;45(12):2279-2294 [FREE Full text] [doi: 10.1080/02664763.2017.1414166]
78. Spicher N, Barakat R, Wang J, Haghi M, Jagieniak J, Öktem GS, et al. Proposing an international standard accident number for interconnecting information and communication technology systems of the rescue chain. Methods Inf Med 2021 Jun;60(S 01):e20-e31 [FREE Full text] [doi: 10.1055/s-0041-1728676] [Medline: 33979848]
79. Haghi M, Barakat R, Spicher N, Heinrich C, Jageniak J, Öktem GS, et al. Automatic information exchange in the early rescue chain using the International Standard Accident Number (ISAN). Healthcare (Basel) 2021 Aug 04;9(8):996 [FREE Full text] [doi: 10.3390/healthcare9080996] [Medline: 34442133]
80. Aceto G, Persico V, Pescapé A. Industry 4.0 and health: internet of things, big data, and cloud computing for healthcare 4.0. J Ind Inf Integr 2020 Jun;18:100129 [FREE Full text] [doi: 10.1016/j.jii.2020.100129]
81. Tamburis O, Masciari E, Fatone G. Development of a decision tree model to improve case detection via information extraction from veterinary electronic medical records. In: Proceedings of the 29th Italian Symposium on Advanced Database Systems. 2021 Presented at: SEBD '21; September 5-9, 2021; Pizzo Calabro, Italy URL: http://ceur-ws.org/Vol-2994/paper10.pdf
82. Anzanpour A, Rahmani AM, Liljeberg P, Tenhunen H. Internet of things enabled in-home health monitoring system using early warning score. In: Proceedings of the 5th EAI International Conference on Wireless Mobile Communication and

Healthcare - Transforming Healthcare through Innovations in Mobile and Wireless Technologies. 2015 Presented at: ICST '15; October 14-16, 2015; London, UK URL: http://kth.diva-portal.org/smash/record.jsf?pid=diva2%3A1471078&dswid=1406 [doi: 10.4108/eai.14-10-2015.2261616]
83. Gerry S, Bonnici T, Birks J, Kirtley S, Virdee PS, Watkinson PJ, et al. Early warning scores for detecting deterioration in adult hospital patients: systematic review and critical appraisal of methodology. BMJ 2020 May 20;369:m1501 [FREE Full text] [doi: 10.1136/bmj.m1501] [Medline: 32434791]
84. Martin V, De Simone L, Lubroth J, Ceccato P, Chevalier V. Perspectives on using remotely-sensed imagery in predictive veterinary epidemiology and global early warning systems. Geospat Health 2007 Nov;2(1):3-14. [doi: 10.4081/gh.2007.250] [Medline: 18686251]
85. Auffray C, Charron D, Hood L. Predictive, preventive, personalized and participatory medicine: back to the future. Genome Med 2010 Aug 26;2(8):57 [FREE Full text] [doi: 10.1186/gm178] [Medline: 20804580]
86. Hood L, Friend SH. Predictive, personalized, preventive, participatory (P4) cancer medicine. Nat Rev Clin Oncol 2011 Mar;8(3):184-187. [doi: 10.1038/nrclinonc.2010.227] [Medline: 21364692]
87. Hood L, Balling R, Auffray C. Revolutionizing medicine in the 21st century through systems approaches. Biotechnol J 2012 Aug;7(8):992-1001 [FREE Full text] [doi: 10.1002/biot.201100306] [Medline: 22815171]
88. Gardes J, Maldivi C, Boisset D, Aubourg T, Vuillerme N, Demongeot J. Maxwell®: an unsupervised learning approach for 5P medicine. Stud Health Technol Inform 2019 Aug 21;264:1464-1465. [doi: 10.3233/SHTI190486] [Medline: 31438183]
89. Benis A, Seidmann A, Ashkenazi S. Reasons for taking the COVID-19 vaccine by US social media users. Vaccines (Basel) 2021 Mar 29;9(4):315 [FREE Full text] [doi: 10.3390/vaccines9040315] [Medline: 33805283]
90. Benis A, Khodos A, Ran S, Levner E, Ashkenazi S. Social media engagement and influenza vaccination during the COVID-19 pandemic: cross-sectional survey study. J Med Internet Res 2021 Mar 16;23(3):e25977 [FREE Full text] [doi: 10.2196/25977] [Medline: 33651709]
91. Syundyukov E, Mednis M, Zaharenko L, Pildegovica E, Danovska I, Kistkins S, et al. Data-driven decision making and proactive citizen-scientist communication: a cross-sectional study on COVID-19 vaccination adherence. Vaccines (Basel) 2021 Nov 24;9(12):1384 [FREE Full text] [doi: 10.3390/vaccines9121384] [Medline: 34960129]
92. Tamburis O, Giannino F, D'Arco M, Tocchi A, Esposito C, Di Fiore G, et al. A night at the OPERA: a conceptual framework for an integrated distributed sensor network-based system to figure out safety protocols for animals under risk of fire. Sensors (Basel) 2020 Apr 29;20(9):2538 [FREE Full text] [doi: 10.3390/s20092538] [Medline: 32365698]
93. Velmovitsky PE, Bevilacqua T, Alencar P, Cowan D, Morita PP. Convergence of precision medicine and public health into precision public health: toward a big data perspective. Front Public Health 2021 Apr 06;9:561873 [FREE Full text] [doi: 10.3389/fpubh.2021.561873] [Medline: 33889555]
94. Eitzel MV, Cappadonna JL, Santos-Lang C, Duerr RE, Virapongse A, West SE, et al. Citizen science terminology matters: exploring key terms. Citiz Sci 2017 Jun 22;2(1):2 [FREE Full text] [doi: 10.5334/cstp.113]
95. Benis A, Amador Nelke S, Winokur M. Training the next industrial engineers and managers about industry 4.0: a case study about challenges and opportunities in the COVID-19 era. Sensors (Basel) 2021 Apr 21;21(9):2905 [FREE Full text] [doi: 10.3390/s21092905] [Medline: 33919164]
96. Wang J, Spicher N, Warnecke JM, Haghi M, Schwartze J, Deserno TM. Unobtrusive health monitoring in private spaces: the smart home. Sensors (Basel) 2021 Jan 28;21(3):864 [FREE Full text] [doi: 10.3390/s21030864] [Medline: 33525460]
97. Ricci FL, Consorti F, Pecoraro F, Luzi D, Tamburis O. A petri-net-based approach for enhancing clinical reasoning in medical education. IEEE Trans Learn Technol 2022 Apr 1;15(2):167-178 [FREE Full text] [doi: 10.1109/tlt.2022.3157391]
98. Osterhaus AD, Vanlangendonck C, Barbeschi M, Bruschke CJ, Christensen R, Daszak P, et al. Make science evolve into a one health approach to improve health and security: a white paper. One Health Outlook 2020;2(1):6 [FREE Full text] [doi: 10.1186/s42522-019-0009-7] [Medline: 32835168]
99. Espeschit ID, Santana CM, Moreira MA. Public policies and one health in Brazil: the challenge of the disarticulation. Front Public Health 2021 Jun 04;9:644748 [FREE Full text] [doi: 10.3389/fpubh.2021.644748] [Medline: 34150698]
100. Transforming our world: the 2030 agenda for sustainable development. Department of Economic and Social Affairs, United Nations. URL: https://sdgs.un.org/2030agenda [accessed 2022-10-28]
101. The 17 goals. Department of Economic and Social Affairs, Sustainable Development, United Nations. URL: https://sdgs.un.org/goals [accessed 2022-10-28]
102. Scquizzato T, Semeraro F. No more unwitnessed out-of-hospital cardiac arrests in the future thanks to technology. Resuscitation 2022 Jan;170:79-81. [doi: 10.1016/j.resuscitation.2021.11.010] [Medline: 34822935]
103. Benis A, Grosjean J, Billey K, Montanha G, Dornauer V, Cri an-Vida M, et al. Medical informatics and digital health multilingual ontology (MIMO): a tool to improve international collaborations. Int J Med Inform 2022 Nov;167:104860 [FREE Full text] [doi: 10.1016/j.ijmedinf.2022.104860] [Medline: 36084537]
104. Darmoni S, Benis A, Lejeune E, Disson F, Dahamna B, Weber P, et al. Digital health multilingual ontology to index teaching resources. Stud Health Technol Inform 2022 Aug 31;298:19-23. [doi: 10.3233/SHTI220900] [Medline: 36073449]
105. Denecke K, Baudoin CR. A review of artificial intelligence and robotics in transformed health ecosystems. Front Med (Lausanne) 2022 Jul 06;9:795957 [FREE Full text] [doi: 10.3389/fmed.2022.795957] [Medline: 35872767]

106. Solomonides AE, Koski E, Atabaki SM, Weinberg S, McGreevey JD, Kannry JL, et al. Defining AMIA's artificial intelligence principles. J Am Med Inform Assoc 2022 Mar 15;29(4):585-591 [FREE Full text] [doi: 10.1093/jamia/ocac006] [Medline: 35190824]
107. Goodman KW. Ethics in health informatics. Yearb Med Inform 2020 Aug;29(1):26-31 [FREE Full text] [doi: 10.1055/s-0040-1701966] [Medline: 32303095]
108. Seh AH, Zarour M, Alenezi M, Sarkar AK, Agrawal A, Kumar R, et al. Healthcare data breaches: insights and implications. Healthcare (Basel) 2020 May 13;8(2):133 [FREE Full text] [doi: 10.3390/healthcare8020133] [Medline: 32414183]
109. Camboim GF, Zawislak PA, Pufal NA. Driving elements to make cities smarter: evidences from European projects. Technol Forecast Soc Change 2019 May;142:154-167 [FREE Full text] [doi: 10.1016/j.techfore.2018.09.014]
110. Abu-Rayash A, Dincer I. Development of integrated sustainability performance indicators for better management of smart cities. Sustain Cities Soc 2021 Apr;67:102704 [FREE Full text] [doi: 10.1016/j.scs.2020.102704]
111. Shittu OS, Williams ID, Shaw PJ. Global E-waste management: can WEEE make a difference? A review of e-waste trends, legislation, contemporary issues and future challenges. Waste Manag 2021 Feb 01;120:549-563. [doi: 10.1016/j.wasman.2020.10.016] [Medline: 33308953]
112. Hosseini ES, Dervin S, Ganguly P, Dahiya R. Biodegradable materials for sustainable health monitoring devices. ACS Appl Bio Mater 2021 Jan 18;4(1):163-194 [FREE Full text] [doi: 10.1021/acsabm.0c01139] [Medline: 33842859]
113. Marques G, Pitarma R, M. Garcia NM, Pombo N. Internet of things architectures, technologies, applications, challenges, and future directions for enhanced living environments and healthcare systems: a review. Electronics 2019 Sep 24;8(10):1081 [FREE Full text] [doi: 10.3390/electronics8101081]
114. Tzivian L, Sokolovska J, Grike AE, Kalcenaua A, Seidmann A, Benis A, et al. Quantitative and qualitative analysis of the quality of life of type 1 diabetes patients using insulin pumps and of those receiving multiple daily insulin injections. Health Qual Life Outcomes 2022 Aug 01;20(1):120 [FREE Full text] [doi: 10.1186/s12955-022-02029-2] [Medline: 35915454]
115. Ho MT, Mantello P, Ghotbi N, Nguyen MH, Nguyen HK, Vuong QH. Rethinking technological acceptance in the age of emotional AI: surveying gen Z (Zoomer) attitudes toward non-conscious data collection. Technol Soc 2022 Aug;70:102011 [FREE Full text] [doi: 10.1016/j.techsoc.2022.102011]
116. Kumar N, Acharya D, Lohani D. An IoT-based vehicle accident detection and classification system using sensor fusion. IEEE Internet Things J 2021 Jan 15;8(2):869-880 [FREE Full text] [doi: 10.1109/jiot.2020.3008896]
117. Cinque M, Esposito C, Russo S, Tamburis O. Blockchain-empowered decentralised trust management for the internet of vehicles security. Comput Electr Eng 2020 Sep;86:106722 [FREE Full text] [doi: 10.1016/j.compeleceng.2020.106722]
118. Schomakers EM, Lidynia C, Ziefle M. The role of privacy in the acceptance of smart technologies: applying the privacy calculus to technology acceptance. Int J Hum-Comput Interact 2021 Nov 25;38(13):1276-1289 [FREE Full text] [doi: 10.1080/10447318.2021.1994211]
119. Kumar TM, Dahiya B. Smart economy in smart cities. In: Kumar TM, editor. Smart Economy in Smart Cities: International Collaborative Research: Ottawa, St.Louis, Stuttgart, Bologna, Cape Town, Nairobi, Dakar, Lagos, New Delhi, Varanasi, Vijayawada, Kozhikode, Hong Kong. The Gateway, Singapore: Springer; 2016:3-76.
120. Amato A, Amato F, Angrisani L, Barolli L, Bonavolontà F, Neglia G, et al. Artificial intelligence-based early prediction techniques in agri-tech domain. In: Proceedings of the 13th International Conference on Intelligent Networking and Collaborative Systems. 2021 Presented at: INCoS '21; September 1-3, 2021; Taichung, Taiwan p. 42-48. [doi: 10.1007/978-3-030-84910-8_5]
121. Yang K, Hu Y, Qi H. Digital health literacy: bibliometric analysis. J Med Internet Res 2022 Jul 06;24(7):e35816 [FREE Full text] [doi: 10.2196/35816] [Medline: 35793141]
122. Atkin C, Crosby B, Dunn K, Price G, Marston E, Crawford C, PIONEER Data Hub. Perceptions of anonymised data use and awareness of the NHS data opt-out amongst patients, carers and healthcare staff. Res Involv Engagem 2021 Jun 14;7(1):40 [FREE Full text] [doi: 10.1186/s40900-021-00281-2] [Medline: 34127076]
123. Nanni M, Andrienko G, Barabási AL, Boldrini C, Bonchi F, Cattuto C, et al. Give more data, awareness and control to individual citizens, and they will help COVID-19 containment. Ethics Inf Technol 2021 Feb 02;23(Suppl 1):1-6 [FREE Full text] [doi: 10.1007/s10676-020-09572-w] [Medline: 33551673]
124. Rivas Velarde MC, Tsantoulis P, Burton-Jeangros C, Aceti M, Chappuis P, Hurst-Majno S. Citizens' views on sharing their health data: the role of competence, reliability and pursuing the common good. BMC Med Ethics 2021 May 18;22(1):62 [FREE Full text] [doi: 10.1186/s12910-021-00633-3] [Medline: 34006284]
125. Teo CL, Chee ML, Koh KH, Tseng RM, Majithia S, Thakur S, et al. COVID-19 awareness, knowledge and perception towards digital health in an urban multi-ethnic Asian population. Sci Rep 2021 May 24;11(1):10795 [FREE Full text] [doi: 10.1038/s41598-021-90098-6] [Medline: 34031469]
126. Kalkman S, van Delden J, Banerjee A, Tyl B, Mostert M, van Thiel G. Patients' and public views and attitudes towards the sharing of health data for research: a narrative review of the empirical evidence. J Med Ethics 2022 Jan;48(1):3-13 [FREE Full text] [doi: 10.1136/medethics-2019-105651] [Medline: 31719155]
127. Abul-Husn NS, Kenny EE. Personalized medicine and the power of electronic health records. Cell 2019 Mar 21;177(1):58-69 [FREE Full text] [doi: 10.1016/j.cell.2019.02.039] [Medline: 30901549]

128. Lloyd KC, Khanna C, Hendricks W, Trent J, Kotlikoff M. Precision medicine: an opportunity for a paradigm shift in veterinary medicine. J Am Vet Med Assoc 2016 Jan 01;248(1):45-48 [FREE Full text] [doi: 10.2460/javma.248.1.45] [Medline: 26684088]
129. Pang LY, Argyle DJ. Veterinary oncology: biology, big data and precision medicine. Vet J 2016 Jul;213:38-45. [doi: 10.1016/j.tvjl.2016.03.009] [Medline: 27240913]
130. Wang J. Advancing the Service Sector with Evolving Technologies: Techniques and Principles. Hershey, PA, USA: IGI Global; 2012.
131. Bonacci I, Tamburis O. Deploying new perspectives of network organizations for chronic diseases' integrated management. Int J Inf Syst Serv Sect 2010;2(3):13-27 [FREE Full text] [doi: 10.4018/jisss.2010070102]
132. Dworkin G. Stanford encyclopedia of philosophy archive: fall 2020 edition. Center for the Study of Language and Information, Stanford University. 2020. URL: https://plato.stanford.edu/archives/fall2020/entries/paternalism/ [accessed 2022-10-28]

Abbreviations

A&EI: accident and emergency informatics
AHE: adverse health event
ECG: electrocardiogram
EDA: electrodermal activity
FAIR: Findable, Accessible, Interoperable, and Reusable
IoAHT: Internet of Animal Health Things
IoMT: Internet of Medical Things
IoT: Internet of Things
ISAN: International Standard Accident Number
ODH: One Digital Health
PPG: photoplethysmography
QoL: quality of life
WHO: World Health Organization

Edited by C Lovis; submitted 28.10.22; peer-reviewed by N Demeter, H Novak Lauscher; comments to author 09.01.23; revised version received 15.03.23; accepted 18.04.23; published 19.05.23.

Please cite as:
Benis A, Haghi M, Deserno TM, Tamburis O
One Digital Health Intervention for Monitoring Human and Animal Welfare in Smart Cities: Viewpoint and Use Case
JMIR Med Inform 2023;11:e43871
URL: https://medinform.jmir.org/2023/1/e43871
doi:10.2196/43871
PMID:36305540

Viewpoint

The Necessity of Interoperability to Uncover the Full Potential of Digital Health Devices

Julian D Schwab[1]; Silke D Werle[1]; Rolf Hühne[1]; Hannah Spohn[1]; Udo X Kaisers[2]; Hans A Kestler[1]

[1]Ulm University, Ulm, Germany

[2]University Hospital Ulm, Ulm, Germany

Corresponding Author:
Hans A Kestler
Ulm University
Albert-Einstein-Allee 11
Ulm, 89081
Germany
Phone: 49 731 500 24500
Fax: 49 731 50024502
Email: hans.kestler@uni-ulm.de

Abstract

Personalized health care can be optimized by including patient-reported outcomes. Standardized and disease-specific questionnaires have been developed and are routinely used. These patient-reported outcome questionnaires can be simple paper forms given to the patient to fill out with a pen or embedded in digital devices. Regardless of the format used, they provide a snapshot of the patient's feelings and indicate when therapies need to be adjusted. The advantage of digitizing these questionnaires is that they can be automatically analyzed, and patients can be monitored independently of doctor visits. Although the questions of most clinical patient-reported outcome questionnaires follow defined standards and are evaluated by clinical trials, these standards do not exist for data processing. Interoperable data formats and structures would benefit multilingual and cross-study data exchange. Linking questionnaires to standardized terminologies such as the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) and Logical Observation Identifiers, Names, and Codes (LOINC) would improve this interoperability. However, linking clinically validated patient-reported outcome questionnaires to clinical terms available in SNOMED CT or LOINC is not as straightforward as it sounds. Here, we report our approach to link patient-reported outcomes from health applications to SNOMED CT or LOINC codes. We highlight current difficulties in this process and outline ways to minimize them.

(JMIR Med Inform 2023;11:e49301) doi:10.2196/49301

KEYWORDS

semantic terminology; semantic; terminology; terminologies; data linkage; interoperability; data exchange; SNOMED CT; LOINC; eHealth; patient-reported outcome questionnaires; requirement for standards; standard; standards; PRO; PROM; patient reported

Introduction

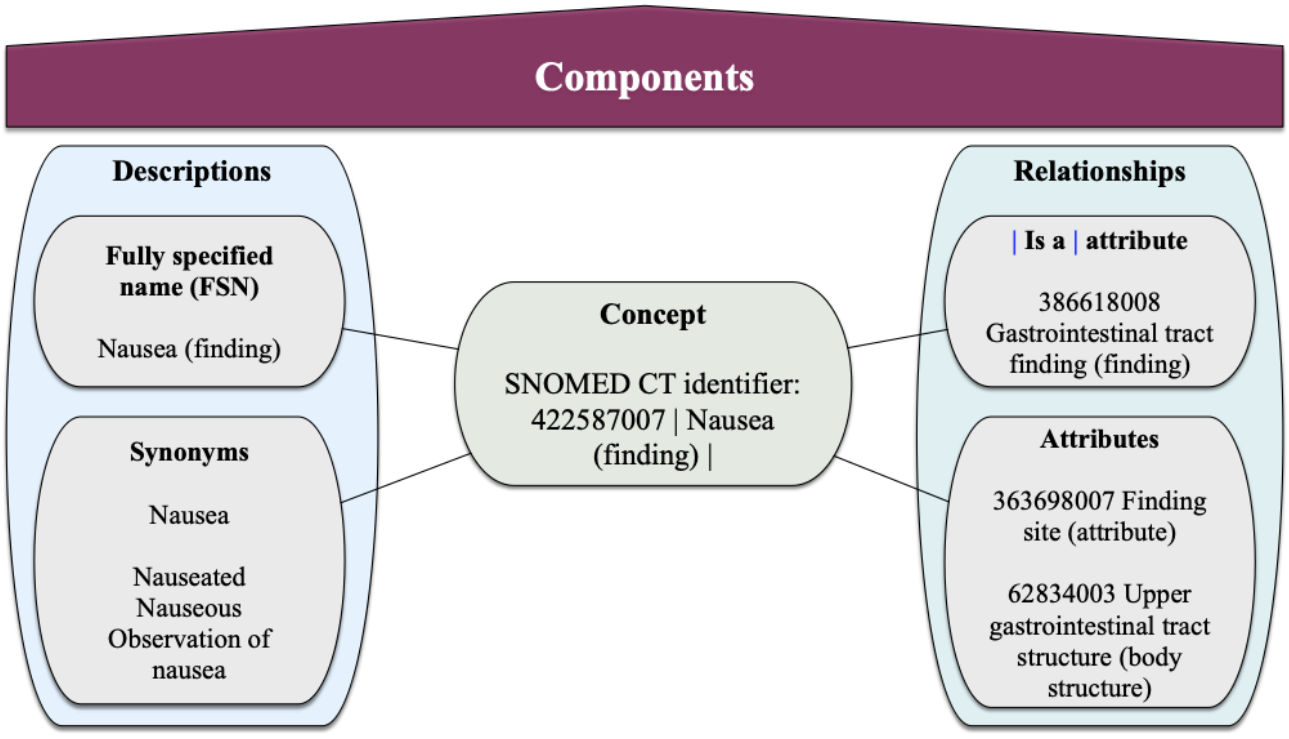
The recording of symptoms and laboratory tests is of paramount importance in clinical practice. In addition to objective, measurable parameters, the patient-reported symptoms and their quality of life have come into focus in modern medicine [1,2]. In this context, patient-reported outcome (PRO) questionnaires have become a standard tool to capture patients' perspectives on their health status. Depending on the underlying disease, PROs cover topics such as quality of life [1], adverse events [3,4], or stress [5], among others. Although in the past, PROs were typically reported using paper-and-pencil methods, electronic PROs (ePROs) are increasingly used and preferred by patients [6]. Because physician and patient perceptions of symptoms can be discrepant [7], PROs are essential for identifying conditions requiring a therapy adjustment [3]. In addition, ePROs can automate their analysis, and they can be used as an outpatient triage tool to identify those in a cohort who need therapy adjustment or closer support [8]. PRO questionnaires can be constructed for different topics and can be more general or tailored to specific diseases [9]. In addition, questions can have different recall time resolutions (eg, today or last week) or differ in their detailed description of symptoms [3]. Thus, the questions and the corresponding answers can be very specific. As a result, minor adjustments in wording can affect the final result [9]. The same is valid for alternative translations. To address these issues, the European Organization for Research and Treatment of Cancer (EORTC), among others, has developed standardized quality-of-life PRO questionnaires

[1]. These questionnaires were initially developed in 1986 and have been refined, translated into several languages, and validated [10,11]. Moreover, the consortium provides a detailed analysis plan to enable comparable results between clinical trials regardless of the geographical region or linguistic and cultural population [10].

Like the underlying questionnaires, the outgoing ePRO data should be accurate, standardized, and interoperable to enable sharing, reuse, and international data comparison. Unfortunately, what seems obvious at first glance is far from reality [12]. The Fast Healthcare Interoperability Resources (FHIR) specification, developed by Health Level Seven International, is the most widely accepted standard for communicating health care data [12]. It aims to provide a comprehensive framework and related standards for exchanging, integrating, sharing, and retrieving electronic health information [13]. The FHIR elements required for PRO questionnaires are Questionnaire and QuestionnaireResponse. A Questionnaire defines a structured set of questions to guide the collection of answers. In addition to the questions, it also defines the answer types and possible answers. It provides detailed control over order, presentation, phraseology, and grouping for consistent data collection. Each Questionnaire requires a QuestionnaireResponse as its counterpart for retrieving and organizing the answers. The QuestionnaireResponse provides a structured set of answers for a specific Questionnaire and must match the definitions of the Questionnaire. Terminologies are integrated as CodeSystems by providing a base URL and the term code. This combines the syntactic interoperability (structure and data format) enabled by standards such as FHIR with the semantic interoperability enabled by health terminologies. The Systematized Nomenclature of Medical Clinical Terms (SNOMED CT) is currently the most appropriate and comprehensive clinical health terminology with natural language properties [14,15]. It includes terminologies for medical concepts, descriptions, and relationships, forming a unique component with a specific identifier [15,16] (Figure 1). Another common medical terminology is the Logical Observation Identifiers Names and Codes (LOINC) [12,14]. A LOINC term is defined as the combination of the LOINC code, a unique identifier, and the fully specified name (FSN), which consists of 5 to 6 parts, including the component or analyte, the observed property, the time of the measurement, the type of system, and the scale of the measurement. Where relevant, the measurement method is included as part 6 [17] (Figure 2). Although SNOMED CT seems appropriate as a reference terminology for multilingual semantic interoperability in health care [18], LOINC has the advantage of also providing “display text” that can be used to link questions (“SURVEY_QUEST_TEXT”) [17]. For these reasons, it may be appropriate to supplement SNOMED CT terminology with LOINC, and several approaches for their mapping have been proposed [19,20].

To improve the international reuse and comparison of PROs in clinical trials, we tested the mapping of standardized and validated questionnaires with their corresponding queries to the SNOMED CT and LOINC terminologies. In the following sections, we introduce these PROs, report on our experience mapping them to SNOMED CT and LOINC, highlight current difficulties, and outline some suggestions for minimizing them.

Figure 1. Exemplary composition of SNOMED CT components for nausea. SNOMED CT content consists of concepts, descriptions, and relationships. Each concept is associated with a set of textual descriptions. The descriptions can be grouped into an FSN, which is unique, and several accepted synonyms, allowing users to apply their preferred terms. In addition to the preferred FSN, a synonym can be selected by a language refset. Relationships link concepts within hierarchies. The |Is a| relationship connects concepts within the same hierarchy, whereas the attribute relationship connects relationships in different hierarchies such as finding site, procedure site, or method. SNOMED CT: Systematized Nomenclature of Medicine Clinical Terms.



Patient-Reported Outcome Questionnaires

Content of Patient-Reported Outcome Questionnaires

Various questionnaires focus either on the quality of life and its impact on the underlying disease or on disease- or treatment-related adverse events. Recent approaches have identified conceptual differences between these PRO measures (PROMs) [21]. On the basis of these findings, we have limited our approach to a few clinically relevant and widely accepted questionnaires that differ in their representation of the domains assessed. Thus, our content analysis included the following PROMs, which can be accessed at

- EORTC Quality of Life of Cancer Patients Questionnaire (QLQ-C30) [22]
- EORTC Quality of Life of Breast Cancer Patients Questionnaire (QLQ-BR23) [22]
- Patient-Reported Outcomes Measurement Information System (PROMIS-29) [23]
- Common Terminology Criteria for Adverse Events (CTCAE) [24].

In the following, we describe the analyzed PROMs in more detail.

EORTC QLQ-C30

The EORTC QLQ-C30 is one of the core validated questionnaires of the EORTC Quality of Life Group. It has been optimized in several revisions since its initial release in 1987 and assesses general aspects and symptoms of patients with cancer based on 30 core items [1]. As there is a wide range of tumor diseases, the questionnaire is not considered to be too disease specific. Five questions of the QLQ-C30 do not consider a recall time interval, whereas the remaining questions refer to the recall time interval *last week*. Two different value scales are used for the answers. Most of the questions can be answered with the text *not at all*, *a little*, *quite a bit*, and *very much*, combined with the numerical values 1 to 4. The others have only text for the first and last values (*very poor* and *excellent*), and the numerical values range from 1 to 7. To score the QLQ-C30, the questions are divided into groups to assess functional scales, symptom-related scales, and global health status, using the Likert method of summing scales [10]. The QLQ-C30 is currently available in version 3.0 and is recommended for use in any new study [10]. The QLQ-C30 core questionnaire can be supplemented by modules containing questions about specific tumor sites, symptoms, or treatments [10,25]. All EORTC questionnaires are available in several languages, but the initial version is in English [10,26]. Special care has been taken in their development to ensure that they are universally understandable, regardless of the level of education or cultural background. To ensure these quality standards, each questionnaire undergoes several rounds of revision, including testing with patients from several countries [26]. The current versions of all EORTC questionnaires and their development status can be requested at the website in [22].

EORTC QLQ-BR23

The EORTC QLQ-BR23 questionnaire is the validated add-on module to the QLQ-C30 core questionnaire for patients with breast cancer. It is recommended for general use, but its updated version (QLQ-BR45) has already completed phase IV testing [10], and a new version will soon be available for newly initiated studies. The QLQ-BR23 questionnaire consists of 23 items that address dealing with symptoms, treatment adverse events, body image, sexual functioning, and future outlook [10]. Similar to the QLQ-C30, the items of the QLQ-BR23 can be grouped for scoring based on the available scoring scheme [10]. This time, however, all questions can be answered with the text *not at all*, *a little*, *quite a bit*, and *very much* combined with the numerical values 1 to 4, and the questions cover a recall time interval of either the last week or the last 4 weeks.

PROMIS-29

The PROMIS is an item bank for constructing patient-reported questionnaires in the context of chronic diseases and conditions. It covers the domains of physical, mental, and social health, which are further subdivided into more precise symptom item banks such as physical function (124 items), pain behavior (39 items), or fatigue (95 items), among others, tailored to a broad population of chronic diseases. This allows comparisons between diseases. The system enables exchanging items or minimizing the number of items within a questionnaire without compromising the reliability [27].

Because of the many possible PROMIS questionnaires, we limited our approach to the PROMIS-29 questionnaire available at the website in [23]. It contains 29 questions derived from the 7 PROMIS categories of physical function, anxiety, depression, fatigue, sleep disturbance, ability to participate in social roles and activities, and pain interference, each measured by 4 questions. It also includes a numeric pain intensity scale ranging from 0 (no pain) to 10 (worst pain imaginable). In addition to the pain intensity scale, all other items have a 5-point response scale (eg, 1=never, 2=rarely, 3=sometimes, 4=often, and 5=always). Except for the questions on physical function and ability to participate in social roles and activities, where the time interval is not specified, the requested recall period is the past 7 days.

CTCAE

The CTCAE is the classic scoring system physicians use to classify side effects during cancer treatment. It is used to grade patient-reported symptoms and those observed by clinicians or from laboratory tests [7,28]. Unlike the EORTC and PROMIS questionnaires, the CTCAE questions were not originally designed in a pen-and-paper format to be given to patients. Therefore, clinicians do not rate each question. It is a descriptive terminology for adverse events coupled with a descriptive grading of the event that occurred [24]. The CTCAE presents scores from 1 to 5 with the associated grading (*mild*, *moderate*, *severe*, *life-threatening*, or *dead*). Each grade has a unique textual clinical description to help to assess the correct grading. Because not all adverse events can be classified into 5 classes, some described adverse events have fewer than 5 grades [24]. The current CTCAE scoring (version 5.0) and previous versions

are available at the website in [29]. We have limited our approach to the current CTCAE version 5.0. However, a CTCAE version 6.0 is already in preparation. Following the idea of the CTCAE, the National Cancer Institute (NCI) has developed a Patient-Reported Outcomes Version of the CTCAE (PRO-CTCAE) scoring system. It contains 78 adverse events from the classic CTCAE, which can be queried based on 124 individual questions [2,30]. The PRO-CTCAE scoring system records adverse events according to attributes such as severity (*none*, *mild*, *moderate*, *severe*, *very severe*), frequency (*never*, *rarely*, *occasionally*, *frequently*, *almost constantly*), or interference (*not at all*, *a little bit*, *somewhat*, *quite a bit*, *very much*) and has a general recall time of the last 7 days [30]. Thus, it lacks the detailed textual descriptions of the symptoms that occurred. In contrast, the health application NEMO (German, Nebenwirkungs-Management Onkologie) tried to combine PRO questions with the classic CTCAE scoring descriptions for daily use [3].

Workflow—Finding SNOMED CT and LOINC Codes for PRO Questionnaires

Limitations

SNOMED CT and LOINC are both optimized for health care and laboratory terminologies. Thus, several approaches have been proposed to map SNOMED CT and LOINC codes [19,20] due to their close knowledge representation formalisms (see Figures 1 and 2). However, even applying this mapping to data for which both terminologies are specialized resulted in incomplete results and required significant human effort to complete the task [19,20].

Therefore, we also manually mapped the questionnaires with their associated questions and response options to assign them SNOMED CT and LOINC codes. The questionnaires analyzed in our study were obtained from [22-24]. The workflow used is shown in Figure 3 and is described in more detail below. The results of the manual mapping can be found in Tables S1-S4 in Multimedia Appendix 1.

We did not consider the relationship between questions and responses in our analyses of term availability. However, we want to inform the reader that the context between questions and possible responses must also be considered and marked in our workflow for actual use cases.

Figure 2. Exemplary composition of a LOINC term for nausea. A LOINC term is defined as the combination of the LOINC code, a unique identifier, and the FSN, which consists of 5-6 parts, including the component or analyte, the observed property, the time of the measurement, the type of system, the scale of the measurement, and sometimes also the method used. The FSN parts are listed sequentially, separated by ":". "Nausea [Presence]" is one of the additional names, a long common name, which must be unique. LOINC: Logical Observation Identifiers, Names, and Codes.

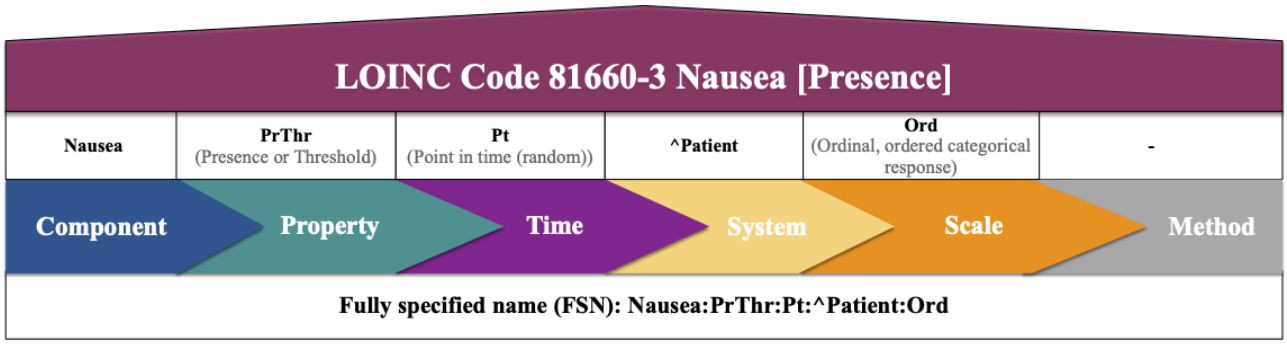
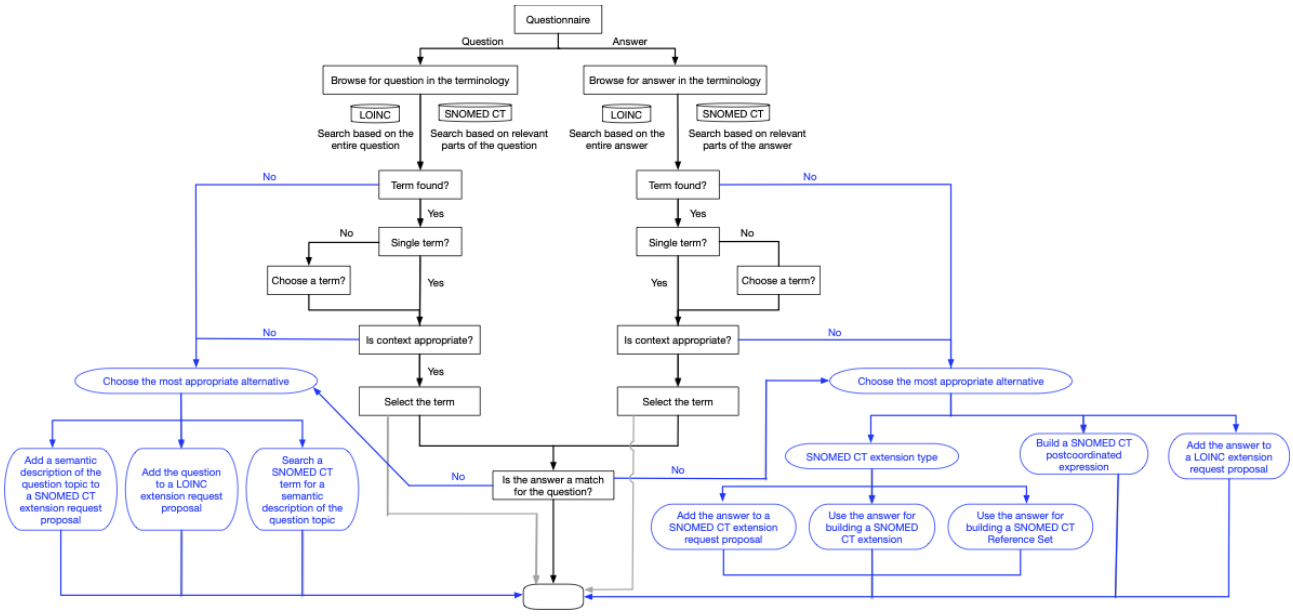


Figure 3. Workflow. Depicted is our workflow to match patient-reported outcome questionnaires with the standardized terminologies LOINC and SNOMET CT to enable semantic interoperability. The workflow marked in black depicts the procedure to analyze the current status of terms included in both terminologies. If a term still needs to be included, we suggest the blue workflow to extend the terminologies. The empty box indicates the end of the procedure. LOINC: Logical Observation Identifiers, Names, and Codes; SNOMED CT: Systematized Nomenclature of Medicine Clinical Terms.



LOINC

After free registration at the website in [31], keywords from our PRO questions and possible answers were queried. If only 1 term was found, it was evaluated if it was a "perfect match." If multiple terms were found, the most appropriate term was selected and considered a perfect match. If this evaluation resulted in a perfect match, the term was chosen as the terminology term. However, if the match was negative or no term was found, the item was treated as "not found." For example, for the question *Have you felt nauseated?* there was no match, and a search for the term *nausea* alone returned 12 results, which are listed as follows:

- *67232-9 | How often did you have nausea in the past 7 days [PhenX]*
- *64713-1 | In which months of the pregnancy did you have frequent nausea or vomiting [PhnX]*
- *81660-3 | Nausea [Presence]*
- *70406-4 | I have nausea in the past 7 days [FACIT]*
- *77711-0 | Patient has anorexia, nausea or vomiting in the past week [UPDRS]*
- *42848-2 | Nausea [CCC]*
- *77510-6 | Can pronounce nausea [AmNART]*
- *28391-1 | Nausea [HIV-SSC]*
- *69711-0 | Did you have nausea or an upset stomach, or the feeling that you were going to have diarrhea [Reported.PHQ]*
- *69682-3 | Bothered by nausea, gas, or indigestion in last 4 weeks [Reported.PHQ]*
- *67231-1 | How often did you have pain in the center of the upper stomach in the past 7 days [PhenX]*
- *96839-6 | Nauseous or had stomach problems when I thought about or was exposed to information about the coronavirus in past 2 weeks*

All of these are not a perfect match to the original question. Therefore, we considered the term not an ideal match.

SNOMED CT

To find SNOMED CT components and the corresponding identifier, we used the SNOMED CT Browser [32] and entered keywords from our PRO questions and possible answers into the search field.

Because full text cannot be entered into SNOMED CT, we searched for relevant parts of the question or answer. For example, for the question *Have you felt nauseated?* the search term *nauseated* or *nausea* returned the following result, which we interpret as a perfect match:

- *422587007 | Nausea (finding) |*

Here, we were interested in a general proof of principle of whether PRO questionnaires can be mapped to semantic terminologies. Therefore, we stopped our analysis after querying the PRO questionnaire terms. However, if the goal is to extend the terminologies, we recommend the following alternatives (marked in blue in Figure 3), which we discuss in more detail below.

In the case of LOINC, the missing term can either be requested to be added by an extension request proposal, or a semantic description of the term can be searched in the SNOMED CT terminology. The specificity of the term description can be further increased by building a postcoordinated expression.

In the following sections, we report on the challenges of term mapping to semantic terminologies.

Challenges in Terminology Binding

Our attempts to complete the exemplary questionnaires' terminology binding revealed the following challenges.

Language

Although the PRO questionnaires are translated into different languages while preserving the original meaning [11], the terminology databases SNOMED CT and LOINC are only available for a limited number of languages [18]. This is counterproductive to the goal of digitization and medical terminologies to improve interoperability, especially if automatic mapping is the ultimate goal. In addition to the international edition of SNOMED CT, there are currently 18 other country-specific editions that cover, at least in part, 10 additional languages: Danish, Dutch, Estonian, Finnish, French, German, Māori, Norwegian, Spanish, and Swedish [32]. However, several European languages are not yet included [18]. On the positive side, at least the same identifiers are used for different languages. However, this should be the case to ensure interoperability. LOINC currently has 20 linguistic variants covering 14 additional languages: Chinese, Dutch, Estonian, French, German, Greek, Italian, Korean, Polish, Portuguese, Russian, Spanish, Turkish, and Ukrainian [33]. Because of the various language alternatives, we limited our mapping to the English versions of the PRO questionnaires and terminology databases, as has been done by others [34]. The language barrier could be reduced by putting more effort into translating the terminologies into more languages. As mentioned above, FHIR, a standard for health data exchange, enables syntactic interoperability [13]. To ensure a correct transition to semantic interoperability, mapping validators, such as the open-source FHIR Validator framework [35], check whether the exported textual questionnaire and the specified identifier match. If the language of the PRO questionnaire and, thus, the survey text from the FHIR export differs from the language of the associated terminology, the program generates a warning. However, the fact that it only generates a warning may allow existing tools to continue to be used.

Concept Availability and Term Selection

The SNOMED CT and LOINC terminologies are comprehensive, but questionnaire questions are often use-case-specific. We limited our approach to searching for general term matches in the terminologies to provide a proof of concept. However, we recognize and want to emphasize that terminology mapping of PRO questionnaires is more than just finding a match for applied use case scenarios. In addition to the term itself, the context must also match. For example, in SNOMED CT, the term *nausea* is a finding that expects a *yes* or *no* response. In contrast, in LOINC, the code *81660-3 | Nausea [Presence]* expects responses such as follows:

- *LA137-2 | None*
- *LA6752-5 | Mild*
- *LA6751-7 | Moderate*
- *LA6750-9 | Severe*
- *LA9041-0 | Resolved*

Therefore, the question and the response must be addressed in the correct context for a proper representation using encoding systems. This also includes the evaluation of multiple isosemantic representations.

In addition to the context, precise wording is paramount for PROs [25]. Mapping the PROMIS-29 questionnaire to LOINC version 2.71 was relatively straightforward because it has its own LOINC profile (*62337-1 | PROMIS item bank - 29 profile*), including codes for questions and response scores. However, caution is still required. Although the LOINC code for the PROMIS-29 panel has basic attributes, including first released and last updated, these version numbers do not match the version number of the underlying questionnaire. The PROMIS-29 questionnaire is available in versions 1.0, 2.0, and 2.1 [36], with changes to the *abilities to participate in social roles and activities* items in versions 2.0 and later. The available LOINC panel, instead, contains only the 1.0 version items. However, the 4 questions that have changed between these versions can also be found in another LOINC panel (*76731-9 | PROMIS short form - ability to participate in social roles and activities 8a - version 2.0*).

In contrast to LOINC, none of the complete PROMIS-29 statements could be mapped to SNOMED CT identifiers. Sometimes, the requested symptom finding could be matched, but without having the full statement (eg, SNOMED CT code: *307077003 | feeling hopeless (finding)*| for the PROMIS-29 statement *In the past seven days I felt hopeless*). However, not all descriptive symptoms (eg, *In the past seven days I felt worthless*) could be found in SNOMED CT. The best SNOMED CT match for the PROMIS-29 response options was obtained with the following:

- The parent concept: *1157335009 | Numeric grade on a scale of 1 to 5 (qualifier)*|
- Its child concept: *1157337001 | Grade 1 on a scale of 1 to 5 (qualifier value)*|
- Up to its child concept: *1157341002 | Grade 5 on a scale of 1 to 5 (qualifier value)*|
- The parent concept: *1157336005 | Numeric grade on a scale of 0 to 10 (qualifier value)*|

Another difficulty is that not all answer options in the PROMIS-29 questionnaire are associated with an ascending scale from 1 to 5. Nevertheless, this scale is sometimes reversed, even though the same text has been assigned. This demonstrates the disadvantage of using only numerical scales alone as terminological concepts for PRO comparison. Instead, it is essential to couple the numerical scales with a semantic representation of their meaning.

For the EORTC QLQ-C30 and QLQ-BR23 questionnaires, finding LOINC terms that semantically matched the questions was difficult. If at all, we could only find phrases formulated as statements that appeared in questions on the EORTC questionnaires, were taken from other questionnaires, and approximated the flow of words. For example, the QLQ-C30 question *Did you need to rest?* has similarities to the LOINC code

- *70815-6 | I need to rest during the day*

from the Functional Assessment of Chronic Illness Therapy (FACIT) Measurement System questionnaire, which is included in LOINC. Sometimes, we also identified multiple LOINC codes that partially matched the searched item from the EORTC

questionnaire. For example, the question *Do you have any trouble taking a long walk?* could be matched to the LOINC code

- *61609-4 | Are you able to go for a walk of at least 15 minutes [PROMIS]* or
- *79006-3 | Are you able to walk more than a mile [PROMIS]*

Overall, we could not match 2 of 30 questions (7%) from the QLQ-C30 questionnaire and 9 of 23 (39%) from the QLQ-BR23 questionnaire. Similar to LOINC, we did not find a single question in the EORTC PROs that exactly matched the SNOMED CT terminology, but we did find at least 1 or more concepts that described the requested symptoms. For example, the QLQ-C30 question *Have you felt nauseated?* could be mapped to the SNOMED CT concept:

- *422587007 | Nausea (finding)|*

Overall, 29 of 30 questions (96%) of the QLQ-C30 and 18 of 23 (78%) of the QLQ-BR23 could be mapped to SNOMED CT terminologies. However, finding appropriate terminologies for response options in SNOMED CT and LOINC has been difficult. The EORTC Quality of Life Group explicitly states that the coupling of textual descriptions and numerical scores is paramount for using their PRO questionnaire concepts [25]. Thus, the simple use of numerical scales such as

- *1157272001 | Numeric grade on a scale of 1 to 4 (qualifier value)|* or

general adjectival modifiers such as

- *89292003 | Rare (qualifier value)|*

would not be sufficient to match the response options. Although the LOINC terminology includes response list codes that include textual descriptions and numerical scores (eg, *LL5215-0*), we could not find a list that correctly matched all scores and associated text.

Adverse events associated with CTCAE-scored drugs have been reported using terminologies such as the Medical Dictionary for Regulatory Activities (MedDRA) [37]. Thus, the CTCAE scoring incorporates certain elements of the MedDRA terminology [24]. In general, there is a mapping between MedDRA and SNOMED CT [37], which can also be used for our purposes. However, this mapping is not freely available. A manual test to find corresponding SNOMED CT identifiers for the first 50 CTCAE terms resulted in 6 missing terms. For the CTCAE response options, the parent concept and its child concepts exist in SNOMED CT:

- Parent concept: *273141005 | Severities (qualifier value)|*
- Child concept: *255604002 | Mild (qualifier value)|*
- Child concept: *6736007 | Moderate (severity modifier) (qualifier value)|*
- Child concept: *2448400 | Severe (severity modifier) (qualifier value)|*
- Child concept: *442452003 | Life threatening severity (qualifier value)|*

However, the last class *dead* of the CTCAE scores is not included in this severity scoring, and there is no direct relationship to adverse events or the meaning of each score for a specific symptom. We have mainly used the European version of SNOMED CT. Nevertheless, we would like to mention that the US version of SNOMED CT has fully equivalent concepts for the CTCAE scores:

- *46411000124101 | Common terminology criteria for adverse events grade 1 (finding)|*
- *446421000124109 | Common terminology criteria for adverse events grade 2 (finding)|*
- *446431000124107 | Common terminology criteria for adverse events grade 3 (finding)|*

However, even with this CTCAE-related scoring, definitions and symptom descriptions are needed for each CTCAE term.

In summary, it is generally challenging to map the exact wording of questions derived from PRO questionnaires to LOINC or SNOMED CT codes if the specific questionnaire is not included in the terminology. However, linking the response options when text is transformed into numeric scales is easier. However, questionnaires like those of the EORTC consortium declare that more is needed to represent numeric response options with the category descriptions [25]. When comparing the 2 ontologies, LOINC was better suited to map PRO questionnaires and their associated response options than SNOMED CT due to its text representation options.

Questionnaire Licenses

Standardized questionnaires are often licensed, for example, from EORTC, and can only be used for patient surveys after obtaining permission. For commercial purposes, permission usually requires payment of a fee. This may prevent their questions and response options from being included in the SNOMED CT or LOINC terminologies. However, even if a licensed version is available, it should be noted that this does not mean that it is freely available for use. For example, LOINC indicates the copyright for terms for which copyright licenses must be considered. Thus, some licensed PRO questionnaires can be found in LOINC, such as the FACIT PRO questionnaire [38] with its 7 panels, including copyright and reference information. This is possible because the English version of FACIT is freely available [38] and can be included in terminology databases. Including at least 1 language version of the PROs in terminology databases would increase the international data exchange and its reusability.

In recent years, there has been a general movement toward “open access” and “open data,” and Creative Commons licenses as an extension of copyright law are seen as a way to protect the origin of data while making it available [39]. In this context, Ehrnsperger and Tietze [40] provide a broad collection of applied cases for patent pledges and open IP. These concepts may also be considered for licensed questionnaires in the future. Furthermore, initiatives such as the Medical Data-Models portal attempt to establish an open metadata repository to overcome the licensing problem [41].

Because standardized questionnaires, such as those of the EORTC Quality of Group, are designed to be culturally independent and have been tested for language comparability [26], linking the data processing to another language would not change the meaning of the data. Instead, the comparability of

the data supports the original intent of the consortia that created each PRO. This reconciliation of different languages would require human effort, but it would be a first step in improving the reusability and comparability of PRO results across studies.

Terminology Changes

Terminologies are not static. LOINC is currently updated twice a year. SNOMED CT International Edition will be updated monthly beginning in 2022. As a result, terms and concepts may become inactive, and definitions may change. Although the meaning of an identifier should be stable, we found LOINC codes assigned to FSNs with several different parts when comparing the 2.64 and 2.75 releases. For example, the LOINC code *18682-5* is mapped to the

- FSN *Ambulance claims attachment:Cmplx:Enctr:^Patient:Set* in the 2.64 release and to the
- FSN *Note:Find:Pt:Ambulance:Doc:Emergency medicine* in the 2.75 release.

So, in the 2.64 release, it was defined as 5 parts. In the current 2.75, it is defined as 6 parts. No part is identical to another.

Time Resolution

Many PRO questionnaires have a general recall period of 7 days [27,42]. However, only some questions in the questionnaire are associated with this period. Some questions do not specify the exact recall period; others include it at the top of a panel but do not specify it for each question. Thus, a semantic matching of the PRO text may not correctly match a textual description that includes the time span, even though both may refer to the same thing. Because the recall time can dilute the remembered details [3], it is essential to include the period in the semantic terminology or at least include it in one of its compositional parts.

Another issue is the semantic representation of time. An example is the textual description of the general PRO recall time, which can be *past week* or *past seven days*. This semantic variation could be resolved using relationships that indicate the same meaning. Notably, some LOINC terms also include the time resolution. Thus, another way to address the issue of including time is to extend the questionnaire FHIR specification in combination with omitting the time range from some LOINC terms. LOINC extension requests could also help to enable full terminology binding of a questionnaire. As a workaround, SNOMED CT concepts could be used to describe the subject of questions semantically.

Opportunities for Improvement

Extension of Terminologies

The interoperability requirement aligns with the original idea of standardized PRO questionnaires that can be used in trials worldwide. However, considering their year of origin, it is evident that digitization aspects were not considered in their development, and most versions were designed in the classical pen-and-paper format. To overcome this limitation, the EORTC Quality of Life Group has developed guidelines for coupling available EORTC instruments with electronic devices [25]. A desirable next step would be to adapt the existing copyright of the PROs to allow at least 1 language version in the terminology databases, as is the case for the FACIT PRO [38].

If licensing issues are not the limiting problem, new codes can be proposed to extend the current databases. In the case of LOINC, the proposal may include new terms for the entire questionnaire, or if some content matches existing terms, they may be included in the submission form and sent to the Regenstrief Institute [17]. Like LOINC, SNOMED CT is also open to requests for additional concepts, coordinated through its members' National Release Centers [16]. Because both terminologies, SNOMED CT and LOINC, are curated databases, it should be expected that not all extension requests will be accepted. In addition, there is a time lag between submitting an extension request and the final release. Special circumstances call for special measures. LOINC allows for advance release of terms for emergency situations (eg, pandemics or new technologies).

However, these prereleases will be reviewed for the next version release and may disappear again. For SNOMED CT, there are 2 other ways to provide additional data: extensions and reference sets [16]. Extensions can be created to support national, local, or organizational needs that may not have international relevance. Instead, reference sets can be made to customize and extend the content for specific needs. However, in addition to the fact that these extensions have limited availability, a disadvantage of these reference sets is that they must be maintained in the future to adapt to new international SNOMED CT releases, which requires more effort than requesting new identifiers.

Combining LOINC and SNOMED CT

Manual mapping of PRO questions has shown a better match using question text from LOINC. If the requested terminology extension is denied or there is too much time between the request and the release, another way to handle questions without matching LOINC terms might be to find SNOMED CT concepts that at least describe the topic of the underlying question. This will not be as specific as a matching question, but it will at least semantically describe the topic of the question. SNOMED CT provides so-called postcoordinated expressions to logically combine multiple identifiers logically to represent a clinical idea at a higher level of detail [16]. A postcoordinated expression can increase the specificity of the topic description for a question. For example, the question *During the past week, have you vomited?* corresponding to the main concept *300359004 | finding of vomiting* is extended by the attribute *Temporal relationship* and the value *per week*, resulting in the following code:

- *300359004 |Finding of vomiting| : 260863009 |Temporal relationship| : 259038000 |per week|*

In the future, it will be possible to use postcoordinated expressions with concrete values such as integer and decimal numbers as attribute values, which are already included in the compositional grammar specification but are not yet available in the International Edition.

Distinction From Other International Medical Terminologies

Here, we have focused on the general applicability of semantic terminologies for PRO questionnaires in the 2 comprehensive terminology systems for clinical settings, SNOMED CT and LOINC, from an international perspective.

However, we would also like to point out some other interesting terminologies that allow PRO mapping, such as the NCI Thesaurus or the MedDRA. While NCI Thesaurus mainly focuses on cancer and thus includes cross-links to CTCAE [43], MedDRA contains a lexicon for adverse events common in clinical trials [34].

A recent study attempted to map free text entries to MedDRA terms manually and found a match for 68% of the terms [34]. However, similar to our findings, the authors encountered the problem that the textual description of adverse events could fit several different terminology codes, but they were mutually exclusive.

Next, we would like to make a small excursion into the general use of these terminologies in the health care sector, where medical documentation is undergoing a paradigm shift. The structured documentation of health data should now facilitate the exchange of patient data and not only be used for statistical and administrative purposes such as billing [44]. This puts the patient at the center of attention [45]. The International Classification of Disease and Related Health Problems (ICD) code is a globally recognized system of medical terminology that provides uniform names for medical diagnoses [46]. Although the semantic terminologies such as SNOMED CT or LOINC are much more fine grained and could, therefore, replace established data collections such as the ICD in health care, they will certainly only be used in the near future in addition to the ICD codes, as the use of ICD codes is legally required [44]. This also requires additional SNOMED CT or LOINC codes that map to the currently accepted ICD code versions ICD-10 or, soon, ICD-11. Because the aim of this viewpoint was to find a reasonably accurate mapping of the PRO questionnaires analyzed, we did not map the PROs studied to the latest ICD-11 code. However, automatic mappings have been proposed for this task [47].

Discussion

Principal Findings

In general, using semantic terminologies is possible to improve the interoperability and reusability of PRO questionnaires and their associated response options. However, this task cannot be automated and requires human effort. In particular, the limited number of available languages limits the general idea of barrier-free interoperability. We have limited our mapping to the English versions of SNOMED CT and LOINC for these reasons. Although this is a limitation of our approach, we believe that this is the most comprehensive version [17,18] and that our results would be even lower for other languages. In addition, the exact mapping of the phrasing needs to be revised. Although the PRO questionnaires examined in SNOMED CT could only be mapped by symptom descriptions, some PRO questions with exact wording could be found in LOINC. Recognizing that we analyzed a limited number of PRO questionnaires, we would also like to highlight the positive development that some have already been entered into LOINC. One reason for not including PRO questionnaires is primarily due to licensing and copyright issues, which currently prohibit the inclusion of the PRO questions in terminologies. Including at least 1 language version of a licensed PRO is desirable. In this context, initiatives such as the MDM-Portal should be mentioned again, which try to provide open formats [41]. Although SNOMED CT offers a wide range of possible numeric response scales, none of the available scales combine numeric scores with textual descriptions required for PROs. Although such a coupling of scores and text is available in LOINC, none of the available scales perfectly matched the response options of the PROs analyzed. In addition, some questions within a questionnaire may have inverted scales, so care must be taken.

Conclusions

In this study, we manually mapped the PRO questionnaires EORTC QLQ-C30 and QLQ-BR23 as well as the PROMIS-29 and a CTCAE-based questionnaire to the 2 widely used standardized terminologies in health care settings, SNOMED CT and LOINC. We showed that among the PRO questionnaires analyzed, only the PROMIS-29 is fully available in LOINC, whereas for the others, between 60.9% and 93.3% of PRO questions could be linked to the LOINC terminology. Although the PROMIS questionnaire is unavailable in SNOMED CT, the American version includes the CTCAE questions, and 78.3%-96.7% of the EORTC questions could be linked to SNOMED CT. Even more critical were our findings concerning the response options. Except for the PROMIS-29 response options in LOINC, which included a score coupled with displayed text, these responses were not available for other questionnaires in any of the terminologies.

It would be desirable to allow at least 1 language version per licensed questionnaire to be included in the terminologies or to use the open formats for future trials. Moreover, special attention should be paid to linking scores and displayed text in the terminologies, as strongly recommended by the original questionnaire settings.

On the basis of our analysis, we recommend LOINC for the future inclusion of additional PRO questionnaires due to its ability to include displayed text.

Acknowledgments

We thank the “Zentrum für Innovative Versorgung” for the fruitful discussion on mapping the analyzed patient-reported outcome questionnaires. HAK acknowledges funding provided by the Ministry of Science and Art Baden-Württemberg (Zentrum für Innovative Versorgung); the German Federal Ministry of Education and Research, as part of the DIFUTURE project (Medical

Informatics Initiative, grant 01ZZ1804I); and the Ministry of Social Affairs and Integration Baden-Württemberg as part of the project feelBack (networked, digital, patient-related psycho-oncology feedback).

Data Availability

All data generated or analyzed during this study will be included in the published article (and its supplementary information files).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Mapping table for translation of questionnaires to SNOMED CT and LOINC.
[XLSX File (Microsoft Excel File), 57 KB - medinform_v11i1e49301_app1.xlsx]

References

1. Aaronson NK, Ahmedzai S, Bergman B, Bullinger M, Cull A, Duez NJ, et al. The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. J Natl Cancer Inst 1993;85(5):365-376. [doi: 10.1093/jnci/85.5.365] [Medline: 8433390]
2. Kluetz PG, Chingos DT, Basch EM, Mitchell SA. Patient-reported outcomes in cancer clinical trials: measuring symptomatic adverse events with the National Cancer Institute's Patient-Reported Outcomes version of the Common Terminology Criteria for Adverse Events (PRO-CTCAE). Am Soc Clin Oncol Educ Book 2016;35:67-73 [FREE Full text] [doi: 10.1200/EDBK_159514] [Medline: 27249687]
3. Kestler AMR, Kühlwein SD, Kraus JM, Schwab JD, Szekely R, Thiam P, et al. Digitalization of adverse event management in oncology to improve treatment outcome—a prospective study protocol. PLoS One 2021;16(6):e0252493 [FREE Full text] [doi: 10.1371/journal.pone.0252493] [Medline: 34086740]
4. Schwab JD, Schobel J, Werle SD, Fürstberger A, Ikonomi N, Szekely R, et al. Perspective on mHealth concepts to ensure users‘ empowerment–from adverse event tracking for COVID-19 vaccinations to oncological treatment. IEEE Access 2021;9:83863-83875 [FREE Full text] [doi: 10.1109/access.2021.3087315]
5. Schobel J, Volz M, Hörner K, Kuhn P, Jobst F, Schwab JD, et al. Supporting medical staff from psycho-oncology with smart mobile devices: insights into the development process and first results. Int J Environ Res Public Health 2021;18(10):5092 [FREE Full text] [doi: 10.3390/ijerph18105092] [Medline: 34064987]
6. Bischoff-Ferrari HA, Vondechend M, Bellamy N, Theiler R. Validation and patient acceptance of a computer touch screen version of the WOMAC 3.1 osteoarthritis index. Ann Rheum Dis 2005;64(1):80-84 [FREE Full text] [doi: 10.1136/ard.2003.019307] [Medline: 15231508]
7. Xiao C, Polomano R, Bruner DW. Comparison between patient-reported and clinician-observed symptoms in oncology. Cancer Nurs 2013;36(6):E1-E16 [FREE Full text] [doi: 10.1097/NCC.0b013e318269040f] [Medline: 23047799]
8. Sivanandan MA, Sharma C, Bullard P, Christian J. Digital patient-reported outcome measures for monitoring of patients on cancer treatment: cross-sectional questionnaire study. JMIR Form Res 2021;5(8):e18502 [FREE Full text] [doi: 10.2196/18502] [Medline: 34398785]
9. Malay S, Chung KC. How to use outcomes questionnaires: pearls and pitfalls. Clin Plast Surg 2013;40(2):261-269 [FREE Full text] [doi: 10.1016/j.cps.2012.10.002] [Medline: 23506766]
10. Fayers PM, Aaronson NK, Bjordal K, Groenvold M, Curran D, Bottomley A. The EORTC QLQ-C30 scoring manual (3rd edition). European Organisation for Research and Treatment of Cancer. Brussels; 2001. URL: https://www.eortc.org/app/uploads/sites/2/2018/02/SCmanual.pdf [accessed 2023-12-07]
11. Wheelwright S, Bjordal K, Bottomley A, Gilberta A, Martinelli F, Pe M, et al. EORTC quality of life group guidelines for developing questionnaire modules (5th edition). European Organisation for Research and Treatment of Cancer. 2021. URL: https://www.eortc.org/app/uploads/sites/2/2022/07/Module-Guidelines-Version-5-FINAL.pdf [accessed 2023-12-07]
12. Lehne M, Sass J, Essenwanger A, Schepers J, Thun S. Why digital medicine depends on interoperability. NPJ Digit Med 2019;2:79 [FREE Full text] [doi: 10.1038/s41746-019-0158-1] [Medline: 31453374]
13. Vorisek CN, Lehne M, Klopfenstein SAI, Mayer PJ, Bartschke A, Haese T, et al. Fast Healthcare Interoperability Resources (FHIR) for interoperability in health research: systematic review. JMIR Med Inform 2022;10(7):e35724 [FREE Full text] [doi: 10.2196/35724] [Medline: 35852842]
14. Bodenreider O, Cornet R, Vreeman DJ. Recent developments in clinical terminologies—SNOMED CT, LOINC, and RxNorm. Yearb Med Inform 2018;27(1):129-139 [FREE Full text] [doi: 10.1055/s-0038-1667077] [Medline: 30157516]
15. Gaudet-Blavignac C, Foufi V, Bjelogrlic M, Lovis C. Use of the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) for processing free text in health care: systematic scoping review. J Med Internet Res 2021;23(1):e24594 [FREE Full text] [doi: 10.2196/24594] [Medline: 33496673]

16. Randorff HA, Kuropatwa R. SNOMED CT starter guide. International Health Terminology Standards Development Organisation. 2014. URL: https://studylib.net/doc/18128896/snomed-ct-starter-guide [accessed 2023-12-07]
17. The Logical Observation Identifiers Names and Codes Committee. Logical Observation Identifiers Names and Codes LOINC users' guide for LOINC version 2.67. Regenstrief Institute. 2019. URL: https://loinc.org/kb/users-guide/ [accessed 2023-12-07]
18. Miñarro-Giménez JA, Cornet R, Jaulent MC, Dewenter H, Thun S, Gøeg KR, et al. Quantitative analysis of manual annotation of clinical text samples. Int J Med Inform 2019;123:37-48 [FREE Full text] [doi: 10.1016/j.ijmedinf.2018.12.011] [Medline: 30654902]
19. Bodenreider O. Issues in mapping LOINC laboratory tests to SNOMED CT. AMIA Annu Symp Proc 2008;2008:51-55 [FREE Full text] [Medline: 18999311]
20. Lee LH, Groß A, Hartung M, Liou DM, Rahm E. A multi-part matching strategy for mapping LOINC with laboratory terminologies. J Am Med Inform Assoc 2014;21(5):792-800 [FREE Full text] [doi: 10.1136/amiajnl-2013-002139] [Medline: 24363318]
21. Rothmund M, Pilz MJ, Egeter N, Lidington E, Piccinin C, Arraras JI, et al. Patient-reported outcome measures for emotional functioning in cancer patients: content comparison of the EORTC CAT core, FACT-G, HADS, SF-36, PRO-CTCAE, and PROMIS instruments. Psychooncology 2023;32(4):628-639 [FREE Full text] [doi: 10.1002/pon.6109] [Medline: 36707461]
22. Questionnaires. EORTC Quality of Life. URL: https://qol.eortc.org/questionnaires/ [accessed 2023-12-07]
23. PROMIS-29 profile v2.0. PROMIS Health Organization and PROMIS Cooperative Group. 2016. URL: https://www.unmc.edu/centric/_documents/PROMIS-29.pdf [accessed 2023-12-07]
24. Common Terminology Criteria for Adverse Events (CTCAE) version 5.0. U.S. Department of Health and Human Services. 2017. URL: https://ctep.cancer.gov/protocoldevelopment/electronic_applications/docs/ctcae_v5_quick_reference_5x7.pdf [accessed 2023-12-07]
25. Kuliś D, Holzner B, Koller M, Ruyskart P, Itani A, Williams P, et al. Guidance on the implementation and management of EORTC quality of life instruments in electronic applications. European Organisation for Research and Treatment of Cancer. Brussels, Belgium; 2018. URL: https://www.eortc.org/app/uploads/sites/2/2018/03/ePRO-guidelines.pdf [accessed 2023-12-07]
26. Kuliś D, Bottomley A, Velikova G, Greimel E, Koller M. EORTC quality of life group translation procedure (4th edition). European Organisation for Research and Treatment of Cancer. 2017. URL: https://www.eortc.org/app/uploads/sites/2/2018/02/translation_manual_2017.pdf [accessed 2023-12-07]
27. Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005-2008. J Clin Epidemiol 2010;63(11):1179-1194 [FREE Full text] [doi: 10.1016/j.jclinepi.2010.04.011] [Medline: 20685078]
28. Di Maio M, Basch E, Bryce J, Perrone F. Patient-reported outcomes in the evaluation of toxicity of anticancer treatments. Nat Rev Clin Oncol 2016;13(5):319-325 [FREE Full text] [doi: 10.1038/nrclinonc.2015.222] [Medline: 26787278]
29. Cancer Therapy Evaluation Program. National Cancer Institute. URL: https://ctep.cancer.gov/protocoldevelopment/electronic_applications/ctc.htm [accessed 2023-12-07]
30. Basch E, Reeve BB, Mitchell SA, Clauser SB, Minasian LM, Dueck AC, et al. Development of the National Cancer Institute's Patient-Reported Outcomes version of the Common Terminology Criteria for Adverse Events (PRO-CTCAE). J Natl Cancer Inst 2014;106(9):dju244 [FREE Full text] [doi: 10.1093/jnci/dju244] [Medline: 25265940]
31. LOINC search. Regenstrief Institute. URL: https://loinc.org/wp-login.php?redirect_to=https%3A%2F%2Floinc.org%2Fsearch%2F&reauth=1 [accessed 2023-12-07]
32. SNOMED CT browser. SNOMED International. URL: https://browser.ihtsdotools.org/? [accessed 2023-12-07]
33. International. Regenstrief Institute. URL: https://loinc.org/international/ [accessed 2023-12-07]
34. Chung AE, Shoenbill K, Mitchell SA, Dueck AC, Schrag D, Bruner DW, et al. Patient free text reporting of symptomatic adverse events in cancer clinical research using the National Cancer Institute's Patient-Reported Outcomes version of the Common Terminology Criteria for Adverse Events (PRO-CTCAE). J Am Med Inform Assoc 2019;26(4):276-285 [FREE Full text] [doi: 10.1093/jamia/ocy169] [Medline: 30840079]
35. team TID. FHIR resource validator. GitHub. URL: https://github.com/inferno-framework/fhir-validator-app [accessed 2023-12-07]
36. Cella D, Choi SW, Condon DM, Schalet B, Hays RD, Rothrock NE, et al. PROMIS adult health profiles: efficient short-form measures of seven health domains. Value Health 2019;22(5):537-544 [FREE Full text] [doi: 10.1016/j.jval.2019.02.004] [Medline: 31104731]
37. Bodenreider O. Using SNOMED CT in combination with MedDRA for reporting signal detection and adverse drug reactions reporting. AMIA Annu Symp Proc 2009;2009:45-49 [FREE Full text] [Medline: 20351820]
38. Webster K, Cella D, Yost K. The Functional Assessment of Chronic Illness Therapy (FACIT) measurement system: properties, applications, and interpretation. Health Qual Life Outcomes 2003;1:79 [FREE Full text] [doi: 10.1186/1477-7525-1-79] [Medline: 14678568]
39. Murray-Rust P. Open data in science. Nat Prec 2008:1-23 [FREE Full text] [doi: 10.1038/npre.2008.1526.1]

40. Ehrnsperger JF, Tietze F. Patent pledges, open IP, or patent pools? Developing taxonomies in the thicket of terminologies. PLoS One 2019;14(8):e0221411 [FREE Full text] [doi: 10.1371/journal.pone.0221411] [Medline: 31430349]
41. Portal for medical data models. Medizinische Fakultät Münster, Universitätsklinikum Münster. URL: https://www.medizin.uni-muenster.de/en/imi/research/digital-health/mdm-portal.html [accessed 2023-12-07]
42. Bennett AV, Dueck AC, Mitchell SA, Mendoza TR, Reeve BB, Atkinson TM, et al. Mode equivalence and acceptability of tablet computer-, interactive voice response system-, and paper-based administration of the U.S. National Cancer Institute's Patient-Reported Outcomes version of the Common Terminology Criteria for Adverse Events (PRO-CTCAE). Health Qual Life Outcomes 2016;14:24 [FREE Full text] [doi: 10.1186/s12955-016-0426-6] [Medline: 26892667]
43. de Coronado S, Wright LW, Fragoso G, Haber MW, Hahn-Dantona EA, Hartel FW, et al. The NCI thesaurus quality assurance life cycle. J Biomed Inform 2009;42(3):530-539 [FREE Full text] [doi: 10.1016/j.jbi.2009.01.003] [Medline: 19475726]
44. Thun S, Dewenter H. ICD-11, ICHI und SNOMED CT—was bedeuten die Systematiken für E-Health-Anwendungen? [ICD-11, ICHI and SNOMED CT—What do the standards mean for eHealth applications?]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2018;61(7):812-820 [in German]. [doi: 10.1007/s00103-018-2759-2] [Medline: 29846742]
45. Hamberger M, Ikonomi N, Schwab JD, Werle SD, Fürstberger A, Kestler AM, et al. Interaction empowerment in mobile health: concepts, challenges, and perspectives. JMIR Mhealth Uhealth 2022;10(4):e32696 [FREE Full text] [doi: 10.2196/32696] [Medline: 35416786]
46. ICD-11 implementation or transition guide. World Health Organization. 2019. URL: https://icd.who.int/en/docs/ICD-11%20Implementation%20or%20Transition%20Guide_v105.pdf [accessed 2023-12-07]
47. Xu J, Fung KW, Bodenreider O. Sequential mapping—a novel approach to map from ICD-10-CM to ICD-11. In: Otero P, Scott P, Martin SZ, editors. MEDINFO 2021: One World, One Health—Global Partnership for Digital Innovation: Proceedings of the 18th World Congress on Medical and Health Informatics. Amsterdam: IOS Press; 2022:96-100.

Abbreviations

CTCAE: Common Terminology Criteria for Adverse Events
EORTC: European Organization for Research and Treatment of Cancer
ePRO: electronic patient-reported outcome
FACIT: Functional Assessment of Chronic Illness Therapy
FHIR: Fast Healthcare Interoperability Resources
FSN: fully specified name
ICD: International Classification of Disease and Related Health Problems
LOINC: Logical Observation Identifiers, Names, and Codes
MedDRA: Medical Dictionary for Regulatory Activities
NCI: National Cancer Institute
NEMO: Nebenwirkungs-Management Onkologie (German)
PRO: patient-reported outcome
PRO-CTCAE: Patient-Reported Outcomes Version of the CTCAE
PROM: patient-reported outcome measure
PROMIS: Patient-Reported Outcomes Measurement Information System
QLQ-BR23: Quality of Life of Breast Cancer Patients Questionnaire
QLQ-BR45: Updated version of the Quality of Life of Breast Cancer Patients Questionnaire
QLQ-C30: Quality of Life of Cancer Patients Questionnaire
SNOMED CT: Systematized Nomenclature of Medicine Clinical Terms

Edited by C Lovis; submitted 24.05.23; peer-reviewed by R Cornet, R Pryss, J Schobel, R Leuchter; comments to author 07.08.23; revised version received 27.09.23; accepted 12.11.23; published 22.12.23.

Please cite as:
Schwab JD, Werle SD, Hühne R, Spohn H, Kaisers UX, Kestler HA
The Necessity of Interoperability to Uncover the Full Potential of Digital Health Devices
JMIR Med Inform 2023;11:e49301
URL: https://medinform.jmir.org/2023/1/e49301
doi:10.2196/49301
PMID:38133917

Original Paper

Monitoring the Implementation of Tobacco Cessation Support Tools: Using Novel Electronic Health Record Activity Metrics

Jinying Chen[1,2,3,4], PhD; Sarah L Cutrona[1,3], MD; Ajay Dharod[2,5,6,7], MD; Stephanie C Bunch[8], MPH; Kristie L Foley[1,5], PhD; Brian Ostasiewski[9], BS; Erica R Hale[1,2], MS; Aaron Bridges[9], BS, MBA; Adam Moses[2], MHA; Eric C Donny[10], PhD; Erin L Sutfin[11], PhD; Thomas K Houston[1,2], MD; iDAPT Implementation Science Center for Cancer Control[12]

[1]iDAPT Implementation Science Center for Cancer Control, Wake Forest University School of Medicine, Winston-Salem, NC, United States

[2]Department of Internal Medicine, Wake Forest University School of Medicine, Winston-Salem, NC, United States

[3]Department of Population and Quantitative Health Sciences, University of Massachusetts Chan Medical School, Worcester, MA, United States

[4]Department of Preventive Medicine and Epidemiology, Boston University Chobanian & Avedisian School of Medicine, Boston, MA, United States

[5]Department of Implementation Science, Division of Public Health Sciences, Wake Forest University School of Medicine, Winston-Salem, NC, United States

[6]Wake Forest Center for Healthcare Innovation, Winston-Salem, NC, United States

[7]Wake Forest Center for Biomedical Informatics, Winston-Salem, NC, United States

[8]Center for Health Analytics, Media, and Policy, RTI International, Research Triangle Park, NC, United States

[9]Clinical & Translational Science Institute, Wake Forest University School of Medicine, Winston-Salem, NC, United States

[10]Department of Physiology and Pharmacology, Wake Forest University School of Medicine, Winston-Salem, NC, United States

[11]Department of Social Sciences and Health Policy, Wake Forest University School of Medicine, Winston-Salem, NC, United States

[12]Wake Forest University School of Medicine, Winston-Salem, NC, United States

Corresponding Author:
Jinying Chen, PhD
iDAPT Implementation Science Center for Cancer Control
Wake Forest University School of Medicine
1 Medical Center Blvd
Winston-Salem, NC, 27101
United States
Phone: 1 617 358 5838
Email: jinchen@wakehealth.edu

Abstract

Background: Clinical decision support (CDS) tools in electronic health records (EHRs) are often used as core strategies to support quality improvement programs in the clinical setting. Monitoring the impact (intended and unintended) of these tools is crucial for program evaluation and adaptation. Existing approaches for monitoring typically rely on health care providers' self-reports or direct observation of clinical workflows, which require substantial data collection efforts and are prone to reporting bias.

Objective: This study aims to develop a novel monitoring method leveraging EHR activity data and demonstrate its use in monitoring the CDS tools implemented by a tobacco cessation program sponsored by the National Cancer Institute's Cancer Center Cessation Initiative (C3I).

Methods: We developed EHR-based metrics to monitor the implementation of two CDS tools: (1) a screening alert reminding clinic staff to complete the smoking assessment and (2) a support alert prompting health care providers to discuss support and treatment options, including referral to a cessation clinic. Using EHR activity data, we measured the completion (encounter-level alert completion rate) and burden (the number of times an alert was fired before completion and time spent handling the alert) of the CDS tools. We report metrics tracked for 12 months post implementation, comparing 7 cancer clinics (2 clinics implemented the screening alert and 5 implemented both alerts) within a C3I center, and identify areas to improve alert design and adoption.

Results: The screening alert fired in 5121 encounters during the 12 months post implementation. The encounter-level alert completion rate (clinic staff acknowledged completion of screening in EHR: 0.55; clinic staff completed EHR documentation of screening results: 0.32) remained stable over time but varied considerably across clinics. The support alert fired in 1074 encounters during the 12 months. Providers acted upon (ie, not postponed) the support alert in 87.3% (n=938) of encounters, identified a

patient ready to quit in 12% (n=129) of encounters, and ordered a referral to the cessation clinic in 2% (n=22) of encounters. With respect to alert burden, on average, both alerts fired over 2 times (screening alert: 2.7; support alert: 2.1) before completion; time spent postponing the screening alert was similar to completing (52 vs 53 seconds) the alert, and time spent postponing the support alert was more than completing (67 vs 50 seconds) the alert per encounter. These findings inform four areas where the alert design and use can be improved: (1) improving alert adoption and completion through local adaptation, (2) improving support alert efficacy by additional strategies including training in provider-patient communication, (3) improving the accuracy of tracking for alert completion, and (4) balancing alert efficacy with the burden.

Conclusions: EHR activity metrics were able to monitor the success and burden of tobacco cessation alerts, allowing for a more nuanced understanding of potential trade-offs associated with alert implementation. These metrics can be used to guide implementation adaptation and are scalable across diverse settings.

(JMIR Med Inform 2023;11:e43097) doi:10.2196/43097

KEYWORDS

medical informatics; electronic health records; EHR metrics; alerts; alert burden; tobacco cessation; monitoring; clinical decision support; implementation science; smoking cessation; decision tool

Introduction

Background

Provider-facing computerized clinical decision support (CDS) tools in electronic health records (EHRs) are common digital health interventions supporting health care quality improvement programs [1-6]. Monitoring (ie, continual evaluation) of the impact of these tools is important for program evaluation and may ultimately contribute to implementation success [7,8]. Approaches for evaluating CDS tools largely rely on surveys, qualitative interviews, and data collected through direct observation or audio/video recording [9-12]. These approaches require substantial human effort (from implementation staff and clinical teams) for data collection. Automated methods leveraging EHR activity data offer a promising solution to reduce the data collection burden, but research on these methods is still in the earliest stage.

This study aimed to develop automatic metrics to monitor the implementation of EHR-embedded CDS tools and demonstrate their use within the context of a smoking cessation program sponsored by a National Cancer Institute (NCI)–designated cancer center.

Tobacco Control Programs in NCI Cancer Centers

Tobacco use increases the risk of cancer and leads to poor prognosis after cancer diagnosis [13-16]. Clinical practice guidelines recommend routine screening for tobacco use and referral to evidence-based cessation interventions in patients with cancer [17,18], but this practice is underused [19]. To address this practice gap, the NCI's Beau Biden Cancer Moonshot program launched the Cancer Center Cessation Initiative (C3I) in 2017 to provide funding to NCI-designated cancer centers to implement or enhance their tobacco treatment services [20].

Electronic alerts (e-alerts) are common CDS tools in EHRs, promoting adherence to practice guidelines [2-5,21,22], including tobacco screening and treatment at the point of care [23-25]. This strategy has been adopted by some C3I-funded cancer centers [26,27]. However, effective implementation of alerts into the clinical workflow is nontrivial [28-32]. Monitoring of provider responses to newly implemented alerts can identify barriers to adoption and the burden imposed by the alerts.

Study Objectives

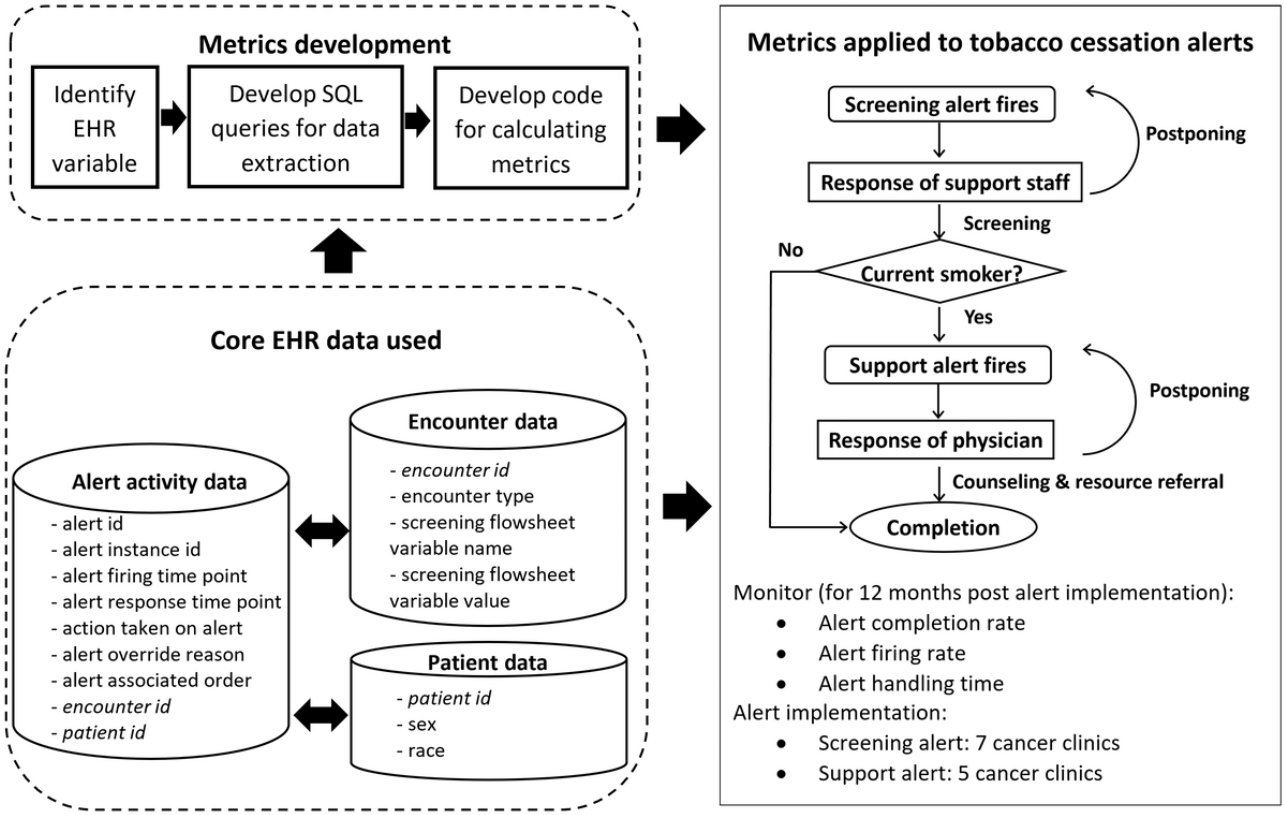
We developed and applied EHR activity metrics to answer three questions. (1) Did the alert completion rate change over time or vary across clinics? (2) What was the burden introduced by the alerts? (3) What factors were associated with variation in alert completion? Our research questions were motivated by three factors. First, sustainability (eg, sustained use and completion of the alerts) is a key construct of implementation outcomes [8] and should be monitored over time. Second, monitoring variations in alert completion across clinics can support the adaptation of alert implementation to the local context. Third, alerts could add a "burden" on providers [30-33], which should be evaluated.

Methods

Study Design

We developed and applied new EHR activity metrics to monitor the tobacco cessation tools (two Best Practice Advisory [BPA] alerts) implemented in cancer clinics for 12 months (Figure 1).

Figure 1. Study overview. The support alert would fire only if the screening result was positive (ie, patient being a current smoker) and answers to both Smoking Screener questions (Q1: "When did you last smoke (even 1 or 2 puffs)?"; Q2: "Quitting smoking could help improve your health. Are you interested in quitting?") were documented (see Multimedia Appendix 1, step B1). EHR: electronic health record.



Ethics Approval

The study was approved by the Wake Forest School of Medicine Institutional Review Board (IRB00066841). Deidentified EHR data were used, with informed consent for data access waved by the institutional review board.

Digital Health Intervention

The CDS tools were two conditional sequential alerts integrated into the Epic EHR, a commercial cloud-based EHR system (Figure 1; detailed in Multimedia Appendix 1): (1) a screening alert to remind clinic staff to complete tobacco screening, triggered if "current smoker" or "unknown smoking status" was previously documented in the EHR, and (2) a support alert to prompt the clinical provider to discuss support and referral to a tobacco cessation clinic, triggered if the screening result was positive and answers to both Smoking Screener questions were answered (see note for Figure 1). Each alert had two modalities: (1) interruptive (triggered when the patient chart was opened; if postponed, presenting again after 10 minutes or when the patient chart was reopened) and (2) noninterruptive (in the general BPA section of the EHR).

Implementation Context

The Tobacco Control Center of Excellence (TCCOE) at the Wake Forest Baptist Comprehensive Cancer Center implemented the alerts in the Epic EHR system used by 7 cancer clinics (medical oncology: n=3; radiation oncology: n=3; cancer survivorship: n=1) in the Atrium Health Wake Forest Baptist Comprehensive Cancer Center in 2019 and 2020. The alerts were integrated into the Epic EHR as BPAs, a form of CDS in the EHR that reminds providers to attend to important tasks [4]. The implementation team from the TCCOE worked with the hospital information technology team on implementing the alerts. The alerts were customized by using rule-based logic (eg, rules on who will receive the alerts and when to fire the alerts; detailed in Multimedia Appendix 1). All 7 clinics implemented the screening alert; 5 implemented the support alert. Training was provided to clinic staff and providers (1-month weekly before or in the first month of implementation and monthly check-in after alert implementation). Some clinics used extensive support from patient navigators and tobacco treatment specialists to complete screening documentation and referral to the cessation clinic.

Evaluation

Metrics Development and Automation

Metrics development took three steps: (1) identifying relevant EHR variables, (2) developing SQL queries to extract variables from the EHR database, and (3) developing computer code to calculate the metrics. We used EHR data associated with 2 clinics to develop and test the metrics. A team of experts in health informatics and implementation science, EHR specialists, and physicians participated in the metrics development.

Using the computer code we developed, EHR data extraction takes about 10 minutes, and the calculation of each metric takes tens of seconds. This speed is adequate for monitoring CDS tools used by implementation programs. Full automation of these metrics is possible after their integration into the EHR.

EHR Variables Used to Derive the Metrics

We extracted alert activity data from event log files of the Epic EHR system. The variables used to develop the metrics included *alert id*, *alert instance id*, *alert name* (eg, a tobacco screening alert), timestamps corresponding to alert firing and provider responding (called *alert firing time point* and *alert response time point* for convenience), *alert triggering condition* (eg, triggered by opening the patient chart), *subsequent actions taken* (eg, acknowledge/override warning), *alert override reason*, and *alert-associated signed orders*.

Each *alert id* is associated with a unique *encounter id* and a *patient id*. An *alert id* corresponds to multiple *alert instance ids* if the alert is fired again after being postponed. *Alert triggering condition* was used to distinguish interruptive alerts from noninterruptive ones.

We used *subsequent actions taken*, *alert override reason*, and *alert-associated signed orders* to identify providers' actions on the alerts. When the clinic staff completed or postponed the screening alert, *subsequent actions taken* recorded a value "acknowledge/override warning," and *alert overridereason* recorded whether the staff acknowledged screening completion (ie, hit the button "Documented in Flowsheet," step A in Figure A1-1, Multimedia Appendix 1), postponed the alert (hit "Defer"), or determined that the patient was inappropriate for screening (hit "Not appropriate"). For the support alert, *subsequent actions taken* recorded a value "acknowledge/override warning" when the provider hit the buttons under "acknowledge reason," and *alert override reason* recorded the provider's actions (eg, discussed or not discussed with patients) and patient's readiness to quit (Figure A1-2, Multimedia Appendix 1). *Alert-associated signed orders* recorded whether the provider placed an order for a referral to the cessation clinic.

In addition, we used two encounter-level variables, *flowsheet name* and *flowsheet value*, to determine whether the clinic staff documented screening results (ie, answers to Q1-Q3 in step B1 in Figure A1-1, Multimedia Appendix 1) in the EHR.

Metrics

We defined three metrics to measure alert completion and burden (Multimedia Appendix 2).

The *alert completion rate* was defined as the number of encounters where a provider completed alert-prompted actions divided by the number of encounters where the alert fired. We defined screening alert completion by either staff's acknowledging completion of screening or completion of EHR documentation of screening results. We defined support alert completion at two levels: (1) discussing with patients and assessing patient readiness to quit ("discussion") or (2) referring patients to the on-site tobacco cessation clinic ("referral").

We measured the burden of interruptive alerts by two metrics: *alert firing rate* and *alert handling time*. We focused on interruptive alerts because they were more likely to add a "burden" on providers [30-33]. We defined *alert firing rate* as the number of times the alert fired during a specific period divided by the number of times the alert was completed during that period. We calculated the average time providers spent completing an alert per encounter, using encounters in which the alert was completed; similarly, we calculated the average time spent postponing alerts per encounter using encounters in which the alert was postponed at least once.

Data Collection

For each clinic, we collected EHR data about clinic characteristics, patient characteristics, and EHR activities related to tobacco cessation alerts. We collected EHR alert activity data as described previously. Each instance of an alert was linked to a specific patient encounter and the patient's demographic information (sex and race), using encounter and patient IDs.

Data Analysis

We summarized clinic and patient characteristics for each clinic. We then used EHR activity metrics to address the research questions related to alert completion and burden. Statistical analyses were conducted using STATA/MP 15.1 (StataCorp LLC) [34].

We measured the overall and per-clinic alert completion rates for the screening alert during every 3-month period across 12 months post alert implementation. The 12-month postimplementation period was specified for each clinic. We measured the support alert completion rate at two levels: "discussion" and "referral."

We measured the alert firing rate and handling time of interruptive screening alerts and support alerts to assess the burden of interruptive alerts.

Factors Associated With Alert Completion

As a secondary analysis, we examined the distribution of alert completion over patients' demographics (sex and race) and encounter types.

Three physicians reviewed all encounter types and selected "relevant encounter" types as those in which screening for smoking status was an appropriate part of routine care (Multimedia Appendix 3).

Results

Clinic Characteristics

The clinics varied in the number of encounters (from n=1464 to n=110,553) and patients (from n=328 to n=9410) during 12 months post alert implementation (Table 1). The typical structure of these clinics was for nurses to support multiple providers across multiple days.

Table 1. Clinic and patient characteristics during the 12 months after implementing tobacco cessation alerts.

	Medical oncology[a]			Radiation oncology[b]			Cancer survivorship (S[c])
	M1[d]	M2	M3	R1[d]	R2	R3	
Clinic characteristics							
Service area	Urban	Rural	Urban	Urban	Rural	Urban	Urban
Staffing, n[e]	5-10	10-20	100-110	10-20	5-10	20-30	10-20
Encounters, n	30,727	9102	110,553	4670	2769	21,362	1464
Patients, n	4688	1193	9410	1059	328	3196	623
Patient characteristics							
Age (years), mean (SD)	64 (14)	65 (13)	61 (15)	66 (11)	67 (11)	64 (14)	59 (19)
Sex, n (%)[f]							
Female	3122 (66.6)	766 (64.2)	4956 (52.7)	603 (56.9)	156 (47.6)	1479 (46.3)	327 (52.5)
Male	1566 (33.4)	427 (35.8)	4454 (47.3)	456 (43.1)	172 (52.4)	1716 (53.7)	296 (47.5)
Race, n (%)[f]							
African American	1070 (22.8)	160 (13.4)	1731 (18.4)	223 (21.1)	50 (15.2)	512 (16.0)	98 (15.7)
White	3350 (71.5)	988 (82.8)	7227 (76.8)	780 (73.7)	263 (80.2)	2546 (79.7)	500 (80.3)
Other[g]	259 (5.5)	45 (3.8)	436 (4.6)	53 (5.0)	14 (4.3)	136 (4.3)	24 (3.9)
Hispanic or Latino, n (%)[f]							
Yes	114 (2.4)	23 (1.9)	349 (3.7)	19 (1.8)	8 (2.4)	85 (2.7)	20 (3.2)
No	4545 (96.9)	1170 (98.1)	9042 (96.1)	1034 (97.6)	320 (97.6)	3104 (97.1)	603 (96.8)
Insurance, n (%)							
Medicare	2721 (58.0)	775 (65.0)	4961 (52.7)	646 (61.0)	205 (62.5)	1699 (53.2)	334 (53.6)
Medicaid	243 (5.2)	81 (6.8)	598 (6.4)	60 (5.7)	19 (5.8)	190 (5.9)	27 (4.3)
Other insurance	1659 (35.4)	305 (25.6)	3537 (37.6)	334 (31.5)	99 (30.2)	1226 (38.4)	253 (40.6)
No insurance	65 (1.4)	32 (2.7)	266 (2.8)	19 (1.8)	5 (1.5)	77 (2.4)	9 (1.4)
Smoking rate, n/N (%)[h]	590/4606 (12.8)	198/1150 (17.2)	1006/9230 (10.9)	174/1051 (16.6)	58/325 (17.8)	435/3155 (13.8)	45/506 (8.9)

[a]M1-M3: medical oncology clinics 1-3.

[b]R1-R3: radiation oncology clinics 1-3.

[c]S: cancer survivorship clinic.

[d]M1 and R1 implemented only the screening alert.

[e]The approximate number of clinic team members (physicians, advanced practice practitioners, nurses, and other clinical staff) in a clinic. The number is not precise due to staff turnover and the hiring of temporary staff.

[f]Some clinics have a small percentage of patients missing information on sex (0.03% missing for R3; complete for other clinics), race (complete for M2; less than 0.3% missing for other clinics), and ethnicity (1% missing for M1, 0.6% missing for R1, 0.2% missing for M3 and R3; complete for other clinics).

[g]Other: American Indian or Alaska Native, Asian, Native Hawaiian or Other Pacific Islander, Latin American or Hispanic, and other.

[h]The percent of patients who were active smokers during 12 months post alert implementation. The denominator is the number of patients who had their smoking status documented in the electronic health record.

Patient Characteristics

The patients seen by the cancer survivorship clinic were 5-8 years older than patients seen by other clinics (mean age for each clinic 59-67; Table 1). Most patients were non-Hispanic White and were beneficiaries of Medicare. The smoking rate ranged between 8.9% (n=45 among 506 patients who had smoking status documented in the EHR; cancer survivorship clinic) and 17.8% (58/325; radiology oncology clinic 2).

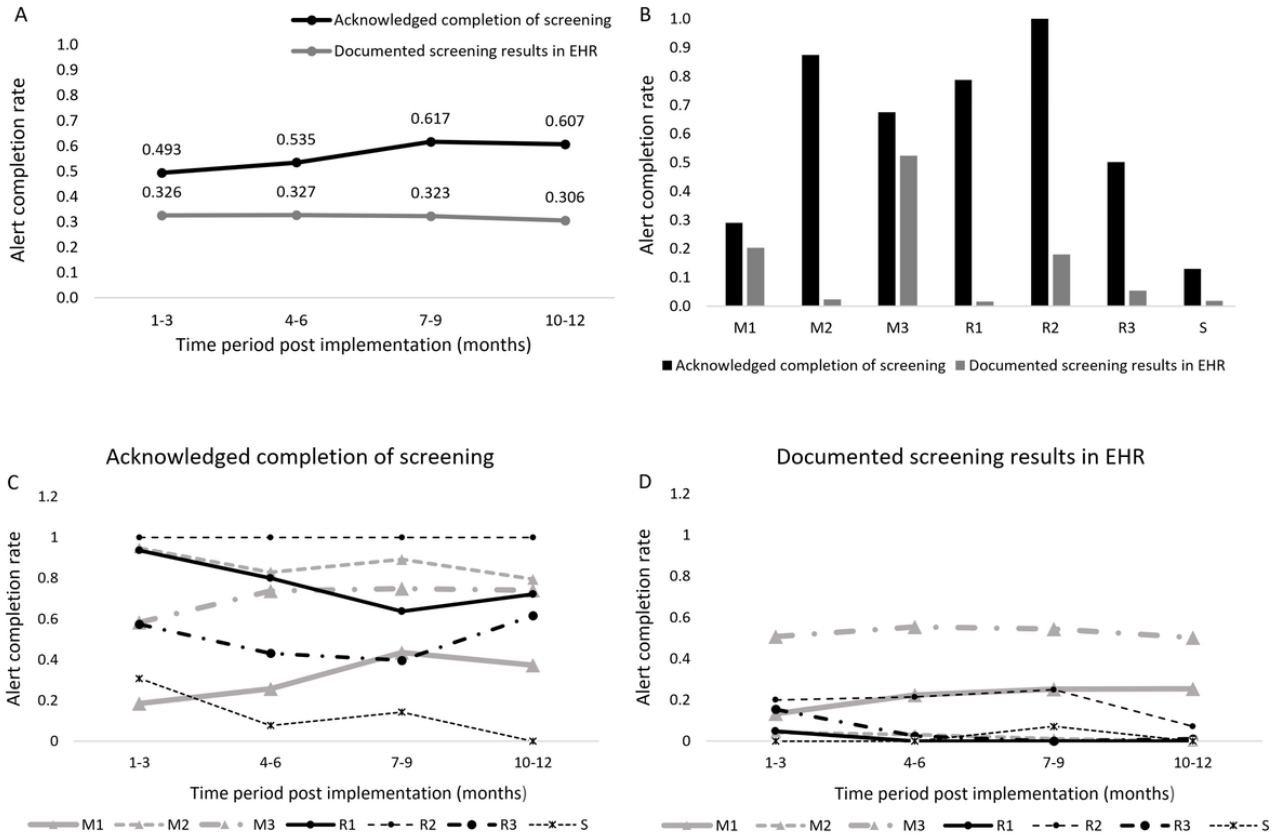
Alert Completion Rate

The screening alert fired in 5121 (2.8% of 180,647) encounters 12 months post implementation. The alert completion rate was 0.55 (2817/5121) based on the staff's acknowledgment of screening completion in EHRs and 0.32 (1647/5121) based on the completion of EHR documentation of screening results. Both alert completion rates remained stable over time (Figure 2A) but varied considerably across clinics (Figure 2B-D).

Among the 2817 encounters where the staff acknowledged completion of screening, 84.7% completed interruptive alerts and 15.4% completed noninterruptive ones.

The support alert was implemented for 5 clinics (medical oncology clinic 2 and 3, radiation oncology clinic 2 and 3, and cancer survivorship clinic) and fired in 1074 encounters. Providers responded without postponing (n=938, 87.3%), discussed tobacco use treatment options (n=640, 59.6%), identified patients who were ready to quit (n=129, 12%), and placed referrals to the cessation clinic (n=22, 2%).

Figure 2. Completion rate of tobacco screening alert for (A) all clinics and (B-D) individual clinics. Clinics in (C) and (D) were categorized into three levels based on the number of encounters in which a screening alert was fired during 12 months post alert implementation. Level 1: >1000; level 2: >100 and ≤1000; level 3: ≤100. Line thickness was used to represent these three levels. EHR: electronic health record. M1-M3: medical oncology clinics 1-3. R1-R3: radiation oncology clinics 1-3. S: cancer survivorship clinic.



The Burden of Interruptive Alerts

On average, the number of times a screening alert was fired before completion was 2.7 (range 1.0-12.7 for individual clinics; Table 2); the average number of times a support alert was fired before completion was 2.1 (range 1.8-3.3 for individual clinics; Table 2).

Table 2. Alert firing rate of the screening alert and the support alert by clinics.

	Medical oncology[a]			Radiation oncology[b]			Cancer survivorship (S[c])
	M1[d]	M2	M3	R1[d]	R2	R3	
Screening alert	4.9[e]	1.3	2.2	2.1	1.0	4.0	12.7
Support alert	N/A[f]	1.8	3.3	N/A	2.3	3.0	3.0

[a]M1-M3: medical oncology clinics 1-3.

[b]R1-R3: radiation oncology clinics 1-3.

[c]S: cancer survivorship clinic.

[d]M1 and R1 implemented only the screening alert.

[e]We defined the alert firing rate as the number of times the alert fired during 12 months post alert implementation divided by the number of times the alert was completed during the same period. We did not calculate the alert firing rate at the encounter level because it was undefined (ie, division by 0) for encounters that did not complete the alert.

[f]N/A: not applicable.

On average, time spent completing the screening alert per encounter was 53 seconds (50 seconds for support alert); time spent postponing screening alerts per encounter was 52 seconds (67 seconds for support alerts).

Factors Associated With Alert Completion

Completion rates of the screening alert and the support alert were balanced across patient subgroups (sex, race, and their interaction).

Among 5121 encounters for which the screening alert was fired, 4425 (86.4%) were "relevant" and 696 (13.6%) were "less relevant" to routine tobacco screening. The alert completion rate for "relevant" encounters was higher than that for "less relevant" ones (2793/4425, 63.1% vs 24/696, 3.5%; $P<.001$).

Discussion

Principal Results

We developed and applied EHR activity metrics to monitor two tobacco cessation CDS alerts implemented in 7 cancer clinics. Our metrics were able to capture variation in alert completion across clinics, monitor alert efficacy, identify discrepancies between staff-acknowledged screening completion and screening documentation, and provide insights into the balance between alert efficacy and imposed burden. These findings inform four areas where CDS tool design or use can be improved (Table 3), which we discuss below.

Table 3. Key findings from the application of the electronic health record (EHR) activity metrics and implications for clinical decision support (CDS) tool design and use.

Key findings from the application of EHR activity metrics	Implications for CDS tools
Variation in alert completion	Potential for improving alert adoption/completion through local adaptation
• The screening alert completion rates varied substantially across the clinics.	Strategies to support use: • Use clinic-specific strategies to support CDS tool adoption
• The screening alert completion rate was higher for encounters perceived as relevant to routine tobacco screening by physicians.	Strategies to support use: • Consider this factor when promoting the use of tobacco cessation CDS tools among health providers
Limited alert efficacy	Potential for improving support alert efficacy
• Providers responded to most support alerts, but few patients were ready to quit, and referral to the tobacco cessation clinic was rare.	Strategies to support use: • Use additional strategies (eg, patient education, provider training in patient-provider communication) to increase the impact of the CDS tools
Inconsistencies between the acknowledgment of alert completion and documented screening	Potential for improving the accuracy of tracking for alert completion
• EHR documentation of screening results was rare for some clinics, even though their clinic staff acknowledged completion of screening for most encounters.	Design: • Improve CDS tool design to allow accurate tracking of screening completion at the alert level Strategies to support use: • Use metrics that can accurately track screening completion, such as metrics calculated based on the completion of EHR documentation of screening results
Interruptive alerts received more responses but also added burden to providers	Importance of balancing alert efficacy with the burden
• Providers were more responsive to interruptive alerts than noninterruptive ones. • Postponing the interruptive alert did not save providers time compared with completing the alert.	Design: • Increase the time interval between postponing and refiring an interruptive alert • Set a threshold to limit the total number of firing of tobacco cessation alerts during a single encounter

Improving Alert Adoption and Completion Through Local Adaptation

Clinics varied substantially in completing the alert, calling for clinic-specific strategies to improve alert adoption. We also identified a modifiable factor (ie, the alert encounter relevance) that affects alert completion. Our physician coauthors considered certain encounter types (eg, initial consultation and office visit) to be relevant for routine tobacco use screening, while others (eg, lab visit and radiation oncology treatment visit) were deemed less relevant. While existing guidelines recommend repeating the smoking assessment at every encounter [17,18], we found that the completion rate of the screening alert was much lower for "less relevant" encounters, which may appear

to be guideline noncompliance. This finding could be informative for committees that develop tobacco screening and treatment guidelines. Implementation teams that want to enforce the "screening at every encounter" rule may need additional strategies. These could include using provider orientation and local champions to influence the culture surrounding tobacco screening [35].

Improving Support Alert Efficacy

Although providers responded to support alerts frequently, referral to the tobacco cessation clinic was rare. One reason was that few patients were ready to quit at the point of care. Future programs may incorporate additional strategies, such as patient education, provider training in patient-provider communication, and addressing patient-level barriers (eg, barriers associated with health beliefs and socioeconomic factors) [36,37]. Note that the 2% referral rate may underestimate the effect of the support alert because it was calculated based on referrals directly linked to the alert. If tobacco treatment specialists contacted the patients interested in quitting after the patient visits, these follow-up activities would be documented elsewhere without a link to the alert, or if a patient chose other treatment methods (eg, quitline or medications), the alert-driven referral would not happen.

Improving Accuracy of Tracking for Alert Completion

The completion rates of the EHR documentation of screening results were lower for some clinics, even though their clinic staff acknowledged screening completion for most encounters. Through discussion with the team coordinating the tobacco cessation program, we identified one major reason for this gap. In clinics using support from patient navigators to complete screening documentation, the clinic staff were likely to bypass the screening but still acknowledged completion. Therefore, measuring EHR documentation is important for the accurate tracking of alert success. We used encounter-level data for this measurement. Alert-level tracking may be necessary for the future development of targeted strategies (eg, provider-specific training) to improve alert adoption. The alert design can be improved to allow this, for example, by disabling the button for acknowledging the completion of screening until the EHR documentation is completed.

Balancing Alert Efficacy With Burden

Although commonly used, effective integration of e-alerts into the clinical workflow has proven difficult [29-33,38,39]. Medication alerts were frequently overridden by health care providers [29,30,33,40], and providers experienced alert-related burden and fatigue [9,29,31,41]. Our study found that postponing the interruptive alert did not save providers time compared with completing the alert. This was partly due to the refiring of postponed alerts. An overabundance of interruptive alerts in EHRs may lead to frequent "postpone" or "override" actions and user dissatisfaction [31-33]. However, our findings do not support disabling the interruptive alerts, as we found that providers were much more responsive to interruptive alerts than noninterruptive ones. One way to alleviate the alert burden is increasing the time interval between postponing and refiring or setting the maximum number of times (eg, 2 or 3) to fire a tobacco cessation alert during each encounter.

Contribution to Implementation Science Methods

New methods are needed for monitoring implementation, including automated approaches that reduce the data collection burden [7,42]. We contributed to this literature by developing automatic EHR activity metrics for monitoring the implementation of CDS tools. Our approach has three merits. First, automatic metrics are suitable for rapid periodic evaluation of implementation programs. These metrics can identify deviations and variations of CDS use at clinic and provider levels, which may inform the selection of key informants for interviews to identify causes of deviation and variation, and the development of strategies to improve CDS design and use. Second, EHR activity data work "behind the scenes" to capture EHR use behavior without interruptions [43-45]. Metrics built on this data can reduce reporting bias and may minimize Hawthorne effects (ie, participants' engagement with an intervention changes when they are aware of attention from observers) [46]. Third, EHRs have been adopted by most US hospitals [47], and EHR-embedded CDS tools are frequently used to support health care quality improvement [1-6]. The ubiquity of EHRs contributes to the generalizability of our approach.

Our work relates to studies using EHR audit logs (one type of EHR activity data) but is different in methodology. The metrics described in these studies measure EHR use and associated burden (eg, total time on EHR, time spent using the EHR after hours, time spent on chart review per patient per day) [48-52] nonspecific to CDS tools. Using EHR audit logs to measure providers' response to a specific EHR tool is challenging, typically involving manual mapping of low-level actions recorded in the log files to EHR use activities [39,50,53]. We used alert activity data generated by Epic's built-in functions to eliminate manual mapping.

Prior studies on alert burden focused on medication alerts and used alert override rate and alert volume as markers for burden in the context of de-implementation [30-33,40]. To our knowledge, this study is the first to systematically measure the burden of preventive care alerts. Our findings did not support a simple de-implementation approach but call for better local adaptation to balance alert efficacy and burden.

Limitations

This study has several limitations. First, the EHR data we analyzed only contained alert-linked referrals to the tobacco cessation clinic. Our analysis may underestimate the actual effect of the support alert. Second, EHR activity data only capture provider interaction with the EHR and lack information about other clinical activities (eg, discussion with patients, pager ringing) during an encounter. In-depth investigations on clinical workflows and their impact on alert response are needed to better understand the variation of alert completion across clinics.

Conclusions

This study developed EHR activity metrics and demonstrated their use in monitoring the impact of CDS tools implemented

by a C3I-funded implementation program that promotes tobacco cessation in patients with cancer. These metrics can be used to guide implementation adaptation and are scalable and adaptable to other settings that use e-alerts to promote adherence to health practice guidelines.

Acknowledgments

The tobacco cessation alerts were implemented by the Tobacco Control Center of Excellence for clinics at the Wake Forest Baptist Comprehensive Cancer Center. The Wake Forest Baptist Comprehensive Cancer Center was renamed Atrium Health Wake Forest Baptist Comprehensive Cancer Center in 2021.

This work was supported by the National Cancer Institute of the US National Institutes of Health (grants P50CA244693 and P30 CA012197, and CRDF award 66590 through the Cancer Center Cessation Initiative Coordinating Center contract) and the National Heart, Lung, and Blood Institute of the US National Institutes of Health (grant K12HL138049 to JC through the Massachusetts Consortium for Cardiopulmonary Implementation Science Scholars K12 Training Program). The study funder had no role in the design of the study; the collection, analysis, and interpretation of the data; and the decision to submit for publication.

Data Availability

Data supporting the study reported in this paper can be made available in deidentified form subject to establishing a data use agreement with the Wake Forest University School of Medicine. The code supporting this study can be accessed on GitHub [54].

Authors' Contributions

All authors take responsibility for the manuscript content, made critical revisions or contributed important intellectual content, and took the decision to submit the final manuscript. KLF, TKH, SLC, JC, ECD, and ELS obtained funding for this study. JC, TKH, and SLC conceptualized the study. JC designed and developed the electronic health record metrics, with advice from TKH and input from all coauthors. AB, AM, BO, SCB, and ERH obtained or provided data. JC, TKH, SLC, and AD analyzed data. JC and TKH visualized and interpreted the results. SLC, AD, SCB, KLF, ECD, and ELS provided critical feedback on metrics design and result interpretation. JC, TKH, SLC, and KLF wrote the first draft of the manuscript. JC, TKH, SLC, AD, SCB, KLF, BO, ERH, AB, AM, ECD, and ELS contributed to the manuscript revision.

Conflicts of Interest

AD serves as an EHR Consultant for the AAMC CORE program. AD is a co-inventor of WHIRL, which is licensed to IllumiCare, Inc. They have an ownership interest in the WHIRL application. AD is also a co-inventor of mPATH. They have equity in Digital Health Navigation (DHN) Solutions, which has licensed mPATH. None of these potential COIs overlap in any way with the content of the current study.

Multimedia Appendix 1
Clinical workflows associated with tobacco cessation alerts.
[PDF File (Adobe PDF File), 801 KB - medinform_v11i1e43097_app1.pdf]

Multimedia Appendix 2
Electronic health record activity metrics.
[PDF File (Adobe PDF File), 57 KB - medinform_v11i1e43097_app2.pdf]

Multimedia Appendix 3
Encounter types relevant to tobacco use screening for patients with cancer.
[PDF File (Adobe PDF File), 13 KB - medinform_v11i1e43097_app3.pdf]

References

1. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med 2020;3:17. [doi: 10.1038/s41746-020-0221-y] [Medline: 32047862]
2. Classification of digital health interventions v1.0: a shared language to describe the uses of digital technology for health. World Health Organization. 2018. URL: https://apps.who.int/iris/handle/10665/260480 [accessed 2023-02-03]
3. McCoy AB, Thomas EJ, Krousel-Wood M, Sittig DF. Clinical decision support alert appropriateness: a review and proposal for improvement. Ochsner J 2014;14(2):195-202 [FREE Full text] [Medline: 24940129]

4. Kawamoto K, Houlihan CA, Balas EA, Lobach DF. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. BMJ 2005 Apr 02;330(7494):765 [FREE Full text] [doi: 10.1136/bmj.38398.500764.8F] [Medline: 15767266]
5. Page N, Baysari MT, Westbrook JI. A systematic review of the effectiveness of interruptive medication prescribing alerts in hospital CPOE systems to change prescriber behavior and improve patient safety. Int J Med Inform 2017 Sep;105:22-30. [doi: 10.1016/j.ijmedinf.2017.05.011] [Medline: 28750908]
6. Kwok R, Dinh M, Dinh D, Chu M. Improving adherence to asthma clinical guidelines and discharge documentation from emergency departments: implementation of a dynamic and integrated electronic decision support system. Emerg Med Australas 2009 Feb;21(1):31-37. [doi: 10.1111/j.1742-6723.2008.01149.x] [Medline: 19254310]
7. Moore GF, Audrey S, Barker M, Bond L, Bonell C, Hardeman W, et al. Process evaluation of complex interventions: Medical Research Council guidance. BMJ 2015 Mar 19;350:h1258 [FREE Full text] [doi: 10.1136/bmj.h1258] [Medline: 25791983]
8. Proctor E, Silmere H, Raghavan R, Hovmand P, Aarons G, Bunger A, et al. Outcomes for implementation research: conceptual distinctions, measurement challenges, and research agenda. Adm Policy Ment Health 2011 Mar;38(2):65-76 [FREE Full text] [doi: 10.1007/s10488-010-0319-7] [Medline: 20957426]
9. Gregory ME, Russo E, Singh H. Electronic health record alert-related workload as a predictor of burnout in primary care providers. Appl Clin Inform 2017 Jul 05;8(3):686-697 [FREE Full text] [doi: 10.4338/ACI-2017-01-RA-0003] [Medline: 28678892]
10. Fraser HSF, Mugisha M, Remera E, Ngenzi JL, Richards J, Santas X, et al. User perceptions and use of an enhanced electronic health record in Rwanda with and without clinical alerts: cross-sectional survey. JMIR Med Inform 2022 May 03;10(5):e32305 [FREE Full text] [doi: 10.2196/32305] [Medline: 35503526]
11. D'Angelo H, Land SR, Mayne RG. Assessing electronic nicotine delivery systems use at NCI-designated cancer centers in the Cancer Moonshot-funded Cancer Center Cessation Initiative. Cancer Prev Res (Phila) 2021 Aug;14(8):763-766 [FREE Full text] [doi: 10.1158/1940-6207.CAPR-21-0105] [Medline: 34127508]
12. Zheng K, Ratwani RM, Adler-Milstein J. Studying workflow and workarounds in electronic health record-supported work to improve health system performance. Ann Intern Med 2020 Jun 02;172(11 Suppl):S116-S122 [FREE Full text] [doi: 10.7326/M19-0871] [Medline: 32479181]
13. National Center for Chronic Disease Prevention and Health Promotion (US) Office on Smoking and Health. The Health Consequences of Smoking—50 Years of Progress: A Report of the Surgeon General. Atlanta, GA: Centers for Disease Control and Prevention; 2014.
14. Toll BA, Brandon TH, Gritz ER, Warren GW, Herbst RS, AACR Subcommittee on Tobacco and Cancer. Assessing tobacco use by cancer patients and facilitating cessation: an American Association for Cancer Research policy statement. Clin Cancer Res 2013 Apr 15;19(8):1941-1948 [FREE Full text] [doi: 10.1158/1078-0432.CCR-13-0666] [Medline: 23570694]
15. Warren GW, Kasza KA, Reid ME, Cummings KM, Marshall JR. Smoking at diagnosis and survival in cancer patients. Int J Cancer 2013 Jan 15;132(2):401-410 [FREE Full text] [doi: 10.1002/ijc.27617] [Medline: 22539012]
16. Balduyck B, Sardari Nia P, Cogen A, Dockx Y, Lauwers P, Hendriks J, et al. The effect of smoking cessation on quality of life after lung cancer surgery. Eur J Cardiothorac Surg 2011 Dec;40(6):1432-7; discussion 1437. [doi: 10.1016/j.ejcts.2011.03.004] [Medline: 21498082]
17. Shields PG, Herbst RS, Arenberg D, Benowitz NL, Bierut L, Luckart JB, et al. Smoking Cessation, Version 1.2016, NCCN Clinical Practice Guidelines in Oncology. J Natl Compr Canc Netw 2016 Nov;14(11):1430-1468. [doi: 10.6004/jnccn.2016.0152] [Medline: 27799513]
18. Hanna N, Mulshine J, Wollins DS, Tyne C, Dresler C. Tobacco cessation and control a decade later: American society of clinical oncology policy statement update. J Clin Oncol 2013 Sep 01;31(25):3147-3157. [doi: 10.1200/JCO.2013.48.8932] [Medline: 23897958]
19. Price SN, Studts JL, Hamann HA. Tobacco use assessment and treatment in cancer patients: a scoping review of oncology care clinician adherence to clinical practice guidelines in the U.S. Oncologist 2019 Feb;24(2):229-238 [FREE Full text] [doi: 10.1634/theoncologist.2018-0246] [Medline: 30446582]
20. Croyle RT, Morgan GD, Fiore MC. Addressing a Core Gap in Cancer Care - The NCI Moonshot Program to Help Oncology Patients Stop Smoking. N Engl J Med 2019 Feb 07;380(6):512-515 [FREE Full text] [doi: 10.1056/NEJMp1813913] [Medline: 30601710]
21. Bernstein SL, Rosner J, DeWitt M, Tetrault J, Hsiao AL, Dziura J, et al. Design and implementation of decision support for tobacco dependence treatment in an inpatient electronic medical record: a randomized trial. Transl Behav Med 2017 Jun;7(2):185-195 [FREE Full text] [doi: 10.1007/s13142-017-0470-8] [Medline: 28194729]
22. Mahabee-Gittens EM, Dexheimer JW, Gordon JS. Development of a tobacco cessation clinical decision support system for pediatric emergency nurses. Comput Inform Nurs 2016 Dec;34(12):560-569 [FREE Full text] [doi: 10.1097/CIN.0000000000000267] [Medline: 27379524]
23. Schindler-Ruwisch JM, Abroms LC, Bernstein SL, Heminger CL. A content analysis of electronic health record (EHR) functionality to support tobacco treatment. Transl Behav Med 2017 Jun;7(2):148-156 [FREE Full text] [doi: 10.1007/s13142-016-0446-0] [Medline: 27800564]

24. Mathias JS, Didwania AK, Baker DW. Impact of an electronic alert and order set on smoking cessation medication prescription. Nicotine Tob Res 2012 Jun;14(6):674-681. [doi: 10.1093/ntr/ntr265] [Medline: 22180576]
25. Boyle R, Solberg L, Fiore M. Use of electronic health records to support smoking cessation. Cochrane Database Syst Rev 2014 Dec 30;2014(12):CD008743 [FREE Full text] [doi: 10.1002/14651858.CD008743.pub3] [Medline: 25547090]
26. Jose T, Ohde JW, Hays JT, Burke MV, Warner DO. Design and pilot implementation of an electronic health record-based system to automatically refer cancer patients to tobacco use treatment. Int J Environ Res Public Health 2020 Jun 06;17(11):4054 [FREE Full text] [doi: 10.3390/ijerph17114054] [Medline: 32517176]
27. Ramsey AT, Chiu A, Baker T, Smock N, Chen J, Lester T, et al. Care-paradigm shift promoting smoking cessation treatment among cancer center patients via a low-burden strategy, Electronic Health Record-Enabled Evidence-Based Smoking Cessation Treatment. Transl Behav Med 2020 Dec 31;10(6):1504-1514 [FREE Full text] [doi: 10.1093/tbm/ibz107] [Medline: 31313808]
28. Khairat S, Marc D, Crosby W, Al Sanousi A. Reasons for physicians not adopting clinical decision support systems: critical analysis. JMIR Med Inform 2018 Apr 18;6(2):e24 [FREE Full text] [doi: 10.2196/medinform.8912] [Medline: 29669706]
29. Légat L, Van Laere S, Nyssen M, Steurbaut S, Dupont AG, Cornu P. Clinical decision support systems for drug allergy checking: systematic review. J Med Internet Res 2018 Sep 07;20(9):e258 [FREE Full text] [doi: 10.2196/jmir.8206] [Medline: 30194058]
30. Poly TN, Islam MM, Yang H, Li YJ. Appropriateness of overridden alerts in computerized physician order entry: systematic review. JMIR Med Inform 2020 Jul 20;8(7):e15653 [FREE Full text] [doi: 10.2196/15653] [Medline: 32706721]
31. McGreevey JD, Mallozzi CP, Perkins RM, Shelov E, Schreiber R. Reducing alert burden in electronic health records: state of the art recommendations from four health systems. Appl Clin Inform 2020 Jan;11(1):1-12 [FREE Full text] [doi: 10.1055/s-0039-3402715] [Medline: 31893559]
32. Chaparro JD, Hussain C, Lee JA, Hehmeyer J, Nguyen M, Hoffman J. Reducing interruptive alert burden using quality improvement methodology. Appl Clin Inform 2020 Jan;11(1):46-58 [FREE Full text] [doi: 10.1055/s-0039-3402757] [Medline: 31940671]
33. Genco EK, Forster JE, Flaten H, Goss F, Heard KJ, Hoppe J, et al. Clinically inconsequential alerts: the characteristics of opioid drug alerts and their utility in preventing adverse drug events in the emergency department. Ann Emerg Med 2016 Feb;67(2):240-248.e3 [FREE Full text] [doi: 10.1016/j.annemergmed.2015.09.020] [Medline: 26553282]
34. Stata Statistical Software: Release 15. Stata. College Station, TX: StataCorp LLC; 2017. URL: https://www.stata.com [accessed 2023-02-03]
35. Powell BJ, Waltz TJ, Chinman MJ, Damschroder LJ, Smith JL, Matthieu MM, et al. A refined compilation of implementation strategies: results from the Expert Recommendations for Implementing Change (ERIC) project. Implement Sci 2015 Feb 12;10:21 [FREE Full text] [doi: 10.1186/s13012-015-0209-1] [Medline: 25889199]
36. Weaver KE, Danhauer SC, Tooze JA, Blackstock AW, Spangler J, Thomas L, et al. Smoking cessation counseling beliefs and behaviors of outpatient oncology providers. Oncologist 2012;17(3):455-462 [FREE Full text] [doi: 10.1634/theoncologist.2011-0350] [Medline: 22334454]
37. Simmons VN, Litvin EB, Patel RD, Jacobsen PB, McCaffrey JC, Bepler G, et al. Patient-provider communication and perspectives on smoking cessation and relapse in the oncology setting. Patient Educ Couns 2009 Dec;77(3):398-403 [FREE Full text] [doi: 10.1016/j.pec.2009.09.024] [Medline: 19846270]
38. Sidebottom AC, Collins B, Winden TJ, Knutson A, Britt HR. Reactions of nurses to the use of electronic health record alert features in an inpatient setting. Comput Inform Nurs 2012 Apr;30(4):218-26; quiz 227. [doi: 10.1097/NCN.0b013e3182343e8f] [Medline: 22045117]
39. Cutrona SL, Fouayzi H, Burns L, Sadasivam RS, Mazor KM, Gurwitz JH, et al. Primary care providers' opening of time-sensitive alerts sent to commercial electronic health record InBaskets. J Gen Intern Med 2017 Nov;32(11):1210-1219 [FREE Full text] [doi: 10.1007/s11606-017-4146-3] [Medline: 28808942]
40. Nanji KC, Slight SP, Seger DL, Cho I, Fiskio JM, Redden LM, et al. Overrides of medication-related clinical decision support alerts in outpatients. J Am Med Inform Assoc 2014;21(3):487-491 [FREE Full text] [doi: 10.1136/amiajnl-2013-001813] [Medline: 24166725]
41. Singh H, Spitzmueller C, Petersen NJ, Sawhney MK, Sittig DF. Information overload and missed test results in electronic health record-based settings. JAMA Intern Med 2013 Apr 22;173(8):702-704 [FREE Full text] [doi: 10.1001/2013.jamainternmed.61] [Medline: 23460235]
42. Willmeroth T, Wesselborg B, Kuske S. Implementation outcomes and indicators as a new challenge in health services research: a systematic scoping review. Inquiry 2019;56:46958019861257 [FREE Full text] [doi: 10.1177/0046958019861257] [Medline: 31347418]
43. Dumais S, Jeffries R, Russell D, Tang D, Teevan J. Understanding user behavior through log data and analysis. In: Olson J, Kellogg W, editors. Ways of Knowing in HCI. New York, NY: Springer; 2014:349-372.
44. Huerta T, Fareed N, Hefner JL, Sieck CJ, Swoboda C, Taylor R, et al. Patient engagement as measured by inpatient portal use: methodology for log file analysis. J Med Internet Res 2019 Mar 25;21(3):e10957 [FREE Full text] [doi: 10.2196/10957] [Medline: 30907733]

45. Wang JK, Ouyang D, Hom J, Chi J, Chen JH. Characterizing electronic health record usage patterns of inpatient medicine residents using event log data. PLoS One 2019;14(2):e0205379 [FREE Full text] [doi: 10.1371/journal.pone.0205379] [Medline: 30726208]
46. Audrey S, Holliday J, Parry-Langdon N, Campbell R. Meeting the challenges of implementing process evaluation within randomized controlled trials: the example of ASSIST (A Stop Smoking in Schools Trial). Health Educ Res 2006 Jun;21(3):366-377 [FREE Full text] [doi: 10.1093/her/cyl029] [Medline: 16740670]
47. Henry J, Pylypchuk Y, Searcy T, Patel V. Adoption of electronic health record systems among U.S. non-federal acute care hospitals: 2008-2015. HealthIT.gov. 2008. URL: https://www.healthit.gov/data/data-briefs/adoption-electronic-health-record-systems-among-us-non-federal-acute-care-1 [accessed 2023-02-03]
48. Rule A, Chiang MF, Hribar MR. Using electronic health record audit logs to study clinical activity: a systematic review of aims, measures, and methods. J Am Med Inform Assoc 2020 Mar 01;27(3):480-490 [FREE Full text] [doi: 10.1093/jamia/ocz196] [Medline: 31750912]
49. Sinsky CA, Rule A, Cohen G, Arndt BG, Shanafelt TD, Sharp CD, et al. Metrics for assessing physician activity using electronic health record log data. J Am Med Inform Assoc 2020 Apr 01;27(4):639-643 [FREE Full text] [doi: 10.1093/jamia/ocz223] [Medline: 32027360]
50. Adler-Milstein J, Adelman JS, Tai-Seale M, Patel VL, Dymek C. EHR audit logs: a new goldmine for health services research? J Biomed Inform 2020 Jan;101:103343 [FREE Full text] [doi: 10.1016/j.jbi.2019.103343] [Medline: 31821887]
51. Lou SS, Lew D, Harford DR, Lu C, Evanoff BA, Duncan JG, et al. Temporal associations between EHR-derived workload, burnout, and errors: a prospective cohort study. J Gen Intern Med 2022 Jul;37(9):2165-2172. [doi: 10.1007/s11606-022-07620-3] [Medline: 35710654]
52. Lou SS, Liu H, Warner BC, Harford D, Lu C, Kannampallil T. Predicting physician burnout using clinical activity logs: model performance and lessons learned. J Biomed Inform 2022 Mar;127:104015. [doi: 10.1016/j.jbi.2022.104015] [Medline: 35134568]
53. Amroze A, Field TS, Fouayzi H, Sundaresan D, Burns L, Garber L, et al. Use of electronic health record access and audit logs to identify physician actions following noninterruptive alert opening: descriptive study. JMIR Med Inform 2019 Feb 07;7(1):e12650 [FREE Full text] [doi: 10.2196/12650] [Medline: 30730293]
54. MIER metrics. GitHub. URL: https://github.com/jchen2017/MIER_metrics [accessed 2023-02-10]

Abbreviations

BPA: Best Practice Advisory
C3I: Cancer Center Cessation Initiative
CDS: clinical decision support
e-alerts: electronic alerts
EHR: electronic health record
NCI: National Cancer Institute
TCCOE: Tobacco Control Center of Excellence

Edited by C Perrin; submitted 29.09.22; peer-reviewed by J Hefner; comments to author 21.10.22; revised version received 21.11.22; accepted 18.01.23; published 02.03.23.

Please cite as:
Chen J, Cutrona SL, Dharod A, Bunch SC, Foley KL, Ostasiewski B, Hale ER, Bridges A, Moses A, Donny EC, Sutfin EL, Houston TK, iDAPT Implementation Science Center for Cancer Control
Monitoring the Implementation of Tobacco Cessation Support Tools: Using Novel Electronic Health Record Activity Metrics
JMIR Med Inform 2023;11:e43097
URL: https://medinform.jmir.org/2023/1/e43097
doi:10.2196/43097
PMID:36862466

Implementation Report

The Journey of Zanzibar's Digitally Enabled Community Health Program to National Scale: Implementation Report

Erica Layer[1], MPH; Salim Slim[2], PhD; Issa Mussa[3], MPH; Abdul-Wahid Al-Mafazy[4], MSc; Giulia V R Besana[3], MSc; Mwinyi Msellem[5], MSc; Isabel Fulcher[1], PhD; Heiko Hornung[6], PhD; Riccardo Lampariello[1], MA

[1]D-tree International, Norwell, MA, United States

[2]Ministry of Health, Zanzibar, United Republic of Tanzania

[3]D-tree International, Zanzibar, United Republic of Tanzania

[4]Office of the Chief Government Statistician, Zanzibar, United Republic of Tanzania

[5]Ministry of Health, Public Health Laboratory, Zanzibar, United Republic of Tanzania

[6]D-tree International, Lusaka, Zambia

Corresponding Author:
Erica Layer, MPH
D-tree International
167 Washington Street
Suite 5
Norwell, MA, 02061
United States
Phone: 1 786314859
Email: elayer@d-tree.org

Abstract

Background: While high-quality primary health care services can meet 80%-90% of health needs over a person's lifetime, this potential is severely hindered in many low-resource countries by a constrained health care system. There is a growing consensus that effectively designed, resourced, and managed community health worker programs are a critical component of a well-functioning primary health system, and digital technology is recognized as an important enabler of health systems transformation.

Objective: In this implementation report, we describe the design and rollout of Zanzibar's national, digitally enabled community health program–Jamii ni Afya.

Methods: Since 2010, D-tree International has partnered with the Ministry of Health Zanzibar to pilot and generate evidence for a digitally enabled community health program, which was formally adopted and scaled nationally by the government in 2018. Community health workers use a mobile app that guides service delivery and data collection for home-based health services, resulting in comprehensive service delivery, access to real-time data, efficient management of resources, and continuous quality improvement.

Results: The Zanzibar government has documented increases in the delivery of health facilities among pregnant women and reductions in stunting among children younger than 5 years since the community health program has scaled. Key success factors included starting with the health challenge and local context rather than the technology, usage of data for decision-making, and extensive collaboration with local and global partners and funders. Lessons learned include the significant time it takes to scale and institutionalize a digital health systems innovation due to the time to generate evidence, change opinions, and build capacity.

Conclusions: Jamii ni Afya represents one of the world's first examples of a nationally scaled digitally enabled community health program. This implementation report outlines key successes and lessons learned, which may have applicability to other governments and partners working to sustainably strengthen primary health systems.

(JMIR Med Inform 2023;11:e48097) doi:10.2196/48097

KEYWORDS

Zanzibar; digital health; community health; health systems strengthening; maternal health; child health; data for decision-making; implementation science; health systems; healthcare infrastructure; health care; implementation report

Introduction

Background

The World Health Organization estimates that high-quality primary health care services can meet 80%-90% of health needs over a person's lifetime [1]. However, in many low-resource countries, this potential is severely hindered by a constrained health care system tackling significant shortages of health workers and the stock of lifesaving medical supplies, all of which limit progress toward universal health coverage (UHC) and damage trust among users [2,3]. There is a growing consensus that effectively designed and managed community health worker (CHW) programs are a critical component of a well-functioning primary health system [4-6]. CHWs have significant potential to extend health services to people's homes, build relationships and meet individual health needs, improve access to services, and improve health system performance [7]. In addition, CHW programs have been shown to deliver a 10:1 return on investment due to increased productivity from a healthier population [8]. Despite this potential, scale-up and management of CHW programs has been slow and often underprioritized by governments [9].

Digital technology has been increasingly recognized as a critical tool to improve both access to health care and the quality of health care delivery, even for CHW programs [10]. Mobile apps, which are designed to fit the local context and are based on program guidelines, can support CHWs to register households and individuals in their communities, thus guiding CHWs through home-based visits following government guidelines. These tools, when integrated into broader digital and public health systems, can improve supervision, coordination, and program monitoring through access to real-time client-level data.

Zanzibar's Community Health Context

The islands of Zanzibar are a semiautonomous region of Tanzania with a population of 1.9 million people. One in 33 women in Zanzibar are estimated to die of maternal complications in her lifetime [11], 20% of children are stunted [12], and Zanzibaris face a high burden of noncommunicable diseases [13]. There is a significant human resource shortage that exacerbates health system challenges; currently, Zanzibar is lacking more than 3900 health workers, which implies a 35% shortage. A key priority for the Ministry of Health Zanzibar is to strengthen its primary health system, including community health [14].

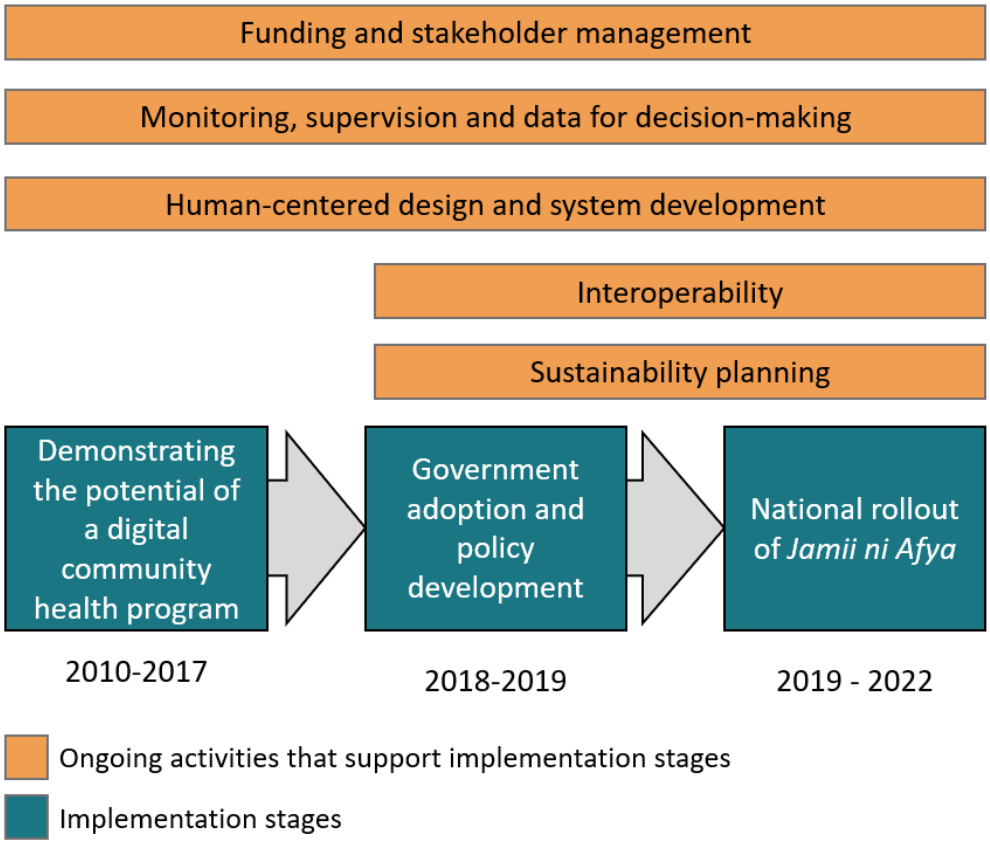
Until recently, community health programs in Zanzibar were traditionally siloed by vertical health programs, funded by donors, and implemented by nongovernmental organizations (NGOs) with programs ending based on donor funding cycles and thematic areas of interest. This resulted in lack of coordination, duplication of efforts, and hindered long-term sustainability. In 2018, the Revolutionary Government of Zanzibar (RGoZ) took the pivotal step of formalizing a government-led community health program, bringing together all previously siloed, donor-funded programs under a single national program, making the pivotal decision to digitize the community health workforce from the start.

In this implementation report, we describe the design and rollout of Zanzibar's national digitally enabled community health program. Our aim is to share lessons learned, drivers for success, and best practices to help other countries succeed in designing, scaling, and sustaining digitally enabled community health programs. This implementation report follows the iCHECK-DH (Guidelines and Checklist for the Reporting on Digital Health Implementations) [15].

Methods

In this section, we present the stages of the evolution of Zanzibar's community health program, which are summarized in Figure 1.

Figure 1. Stages of the evolution of Zanzibar's community health program.



Demonstrating the Potential of a Digital Community Health Program

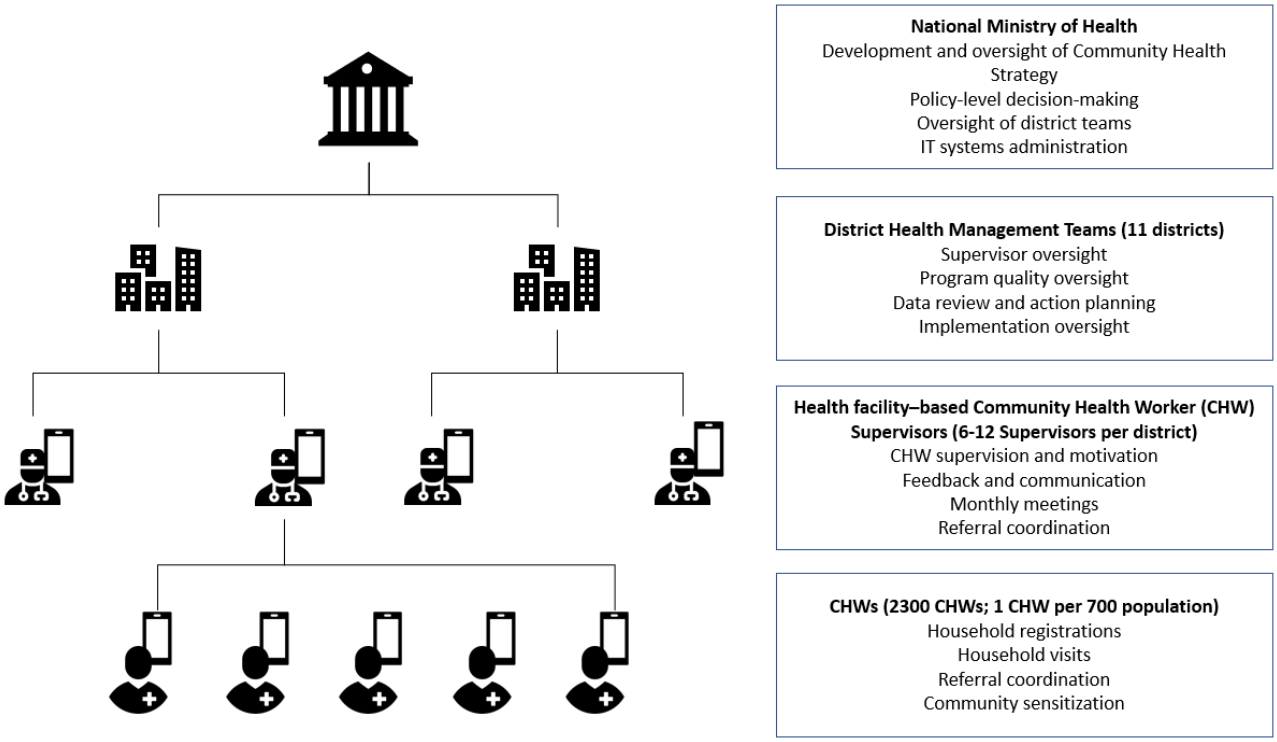
D-tree International, a digital health NGO, began working with the RGoZ in 2010 on Safer Deliveries—a digitally enabled maternal and newborn community health pilot project aimed at increasing facility deliveries and improving postpartum care. This project has been described elsewhere [16,17]. Briefly, CHWs were trained, supervised, and equipped with a mobile app to enable them to register pregnant women, provide health education, screen them for danger signs and refer them to a health facility, and link them with locally available transportation. From 2012 to 2018, this project supported 400 CHWs to care for over 80,000 pregnant women and their newborns and demonstrated an increase in pregnant women delivering at health facilities from 50% to 75% (a 50% increase), led to increases in facility-based postpartum care visits from 20% to 80%, and increased the completion of community facility referrals from 27% to 90% [16]. Clients reported receiving personalized, responsive, and compassionate care. They were empowered to provide direct feedback about the quality of facility-based services, which district health teams used to monitor and improve service quality. Health facility staff and District Health Management Teams were involved in the program design to ensure that health facilities—especially primary health care facilities—were prepared for increased demand for maternal health services.

Government Adoption and Policy Development

Based on the success of the Safer Deliveries project and the recognition that a digitally enabled, government-owned community health system was critical to sustainably improve health in general in Zanzibar, the RGoZ adopted this digital CHW program as part of the formal health system at the national level, unifying all community health programs through this single initiative named "Jamii ni Afya" (Swahili for "Community is Health"). In order to formalize this program as a government initiative, the RGoZ updated its community health strategy to make Jamii ni Afya a central pillar of community health. This strategy serves as a formal government document outlining how community health volunteers fit into the larger Zanzibar health system and specifying their roles, responsibilities, training, and required qualifications. Figure 2 summarizes the structure of the community health program. In particular, the revised community health strategy specifies that CHWs will be supported with a digital health platform, as the Ministry of Health Zanzibar recognized the benefits of technology in the pursuit of its vision for a healthy population.

The program uses the term "community health volunteers" to emphasize engagement on a volunteer basis; however, in line with prior literature, we will henceforth refer to them as CHWs.

Figure 2. Structure of the Zanzibar community health program. CHW: community health worker.



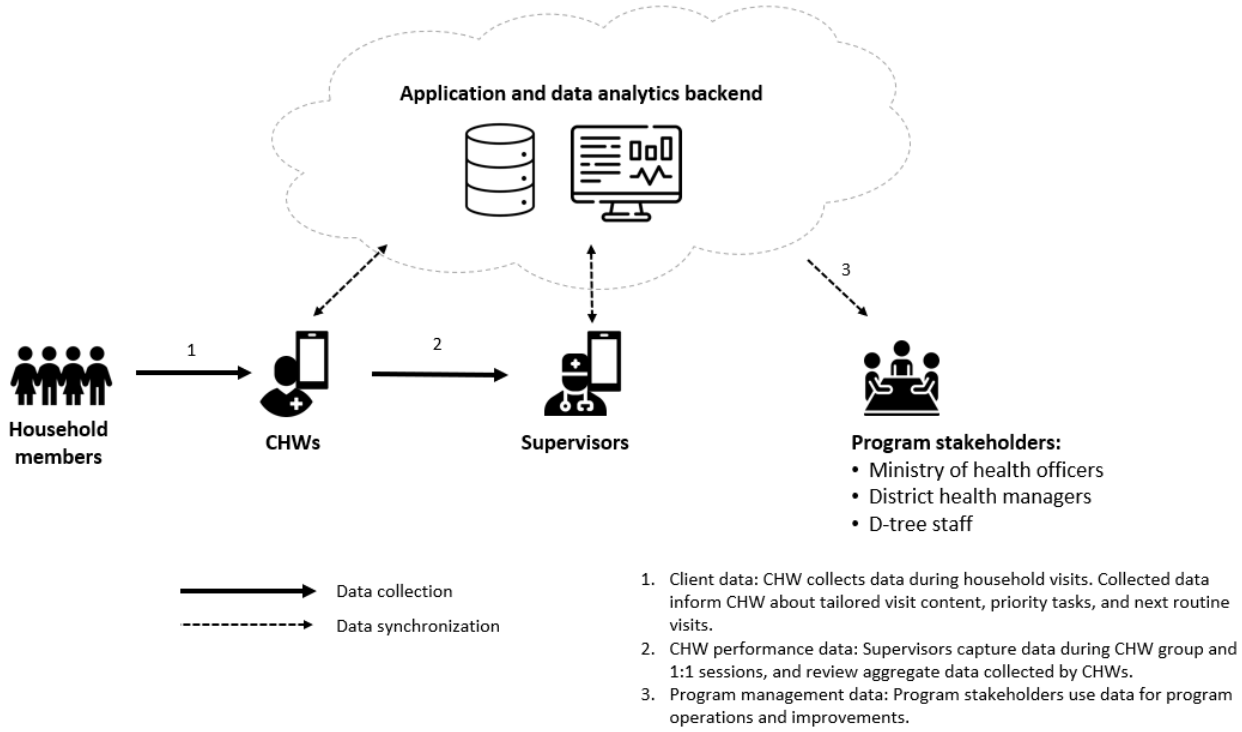
Human-Centered Design and Development

From January to June 2019, the Ministry of Health Zanzibar, D-tree International, and other community health stakeholders participated in a human-centered design process to cocreate both the programmatic and technical aspects of the Jamii ni Afya app. The approach was to fully define the health problem and local context first and then select the technology that is best suited to solve that problem. Thus, the program and technology design processes were conducted simultaneously, where program requirements—including health objectives, CHW workflows, supervisory systems, and long-term government management capacity—informed technology design choices. Community health stakeholders included community members, political and religious village leaders, CHWs, CHW supervisors, and District Health Management Teams. Input was initially gathered through sensitization meetings with all abovementioned community stakeholders and then periodically with district managers. Supervisors also had WhatsApp groups to discuss program questions. During app design and development, CHWs were involved in early prototype testing. Further feedback was gathered during CHW training sessions by observing how well CHWs were able to conduct practical exercises on a training app. After implementation, feedback was gathered during field observations. All activities considered the local context, including gathering buy-in from political village leaders; conducting activities with larger groups and broad participation given the communal culture; and understanding basic digital literacy, resulting in involving CHWs in later app design stages.

The Principles for Digital Development [18] were used throughout the design process, specifically “design with the user,” “understand the existing ecosystem,” “design for scale,” and “reuse and improve” to ensure that the resulting system was locally relevant, user-friendly, cost-effective, and sustainable.

Jamii ni Afya leverages government guidelines and global best practices to guide CHWs using digital technology in delivering high-quality health education and counseling services in maternal and child health, nutrition, water, sanitation and hygiene, and early childhood development. The data generated from these interactions are used to further personalize health services by tailoring health messages based on the personal circumstances of the client, improving supervision, and supporting programmatic and policy decision-making at the community, district, and national levels. For example, if a pregnant woman has specific risk factors (based on government guidelines), the app flags that she should plan to deliver at a hospital rather than a primary health facility, and subsequent counseling within the app supports the woman to develop a birth plan for delivery at the nearest hospital. If a CHW indicates that a child has developmental delays (based on observing the child and caregivers at home), the app suggests specific play activities that the caregivers can undertake to support the child’s development. Figure 3 illustrates the data flow for the program.

Figure 3. Flow of data within Jamii ni Afya. CHW: community health worker.



Interoperability

The Jamii ni Afya mobile app is built on the Community Health Toolkit (CHT), an open-source global goods platform developed to support community health workers globally. This platform was selected by the Zanzibar government due to the following reasons: it is open-source and uses well-known components and frameworks, it has a growing community that can be leveraged for support, it can be hosted in a local data center, the skills required to configure health worker tools are found among Ministry of Information, Communications and Technology staff, and it is easily available in the local market. In addition, CHT runs on low-end Android smartphones and has offline functionality, which is critical in Zanzibar where network connectivity is not guaranteed. Developers from D-tree International and the Ministry of Health Zanzibar led the mobile app development process, including system requirements and specifications. The system was designed to ensure interoperability with the broader digital health ecosystem in Zanzibar. The CHT is integrated with the government's health management information system (District Health Information System [DHIS2]) and provides data for the community health information system. Jamii ni Afya is referenced in the RGoZ's first-ever digital health strategy, which works to coordinate and optimize government resources for digital health in Zanzibar. It is integrated with the Zanzibar Health Interoperability Layer (based on OpenHIM) and can thus integrate with an up and coming electronic medical records system. There are plans to integrate with the future client registry system and a health care worker registry. To date, interoperability has been achieved at the level of data exchange, which is a foundational level focusing on sending and receiving information without interpreting it. Interpretation occurs at the receiving end in accordance with separately defined specifications. For example, when integrating with DHIS2, the submitted data payload was formatted in accordance with DHIS2 requirements, and the meaning of the data payload was agreed on by the involved technical teams. The OpenHIM-based interoperability layer enables higher levels of interoperability, including interoperability of data structures, meaning of data, process, organizational rules, etc. This affords the interpretation of data and information without a system having to understand the DHIS2 or CHT, and makes interoperability with additional systems and platforms easier. However, the technical teams agreed that OpenHIM would initially only pass through payloads—that is, it would not use the higher-level interoperability options—to speed up implementation. For new integrations and for updated integrations, it is planned to leverage terminology and other data standards, which will be consistent with the planned electronic medical record system, but the specific standards are still under discussion.

National Rollout

Jamii ni Afya was introduced in phases from July 2019 to August 2021 to all of Zanzibar's 11 districts. Throughout the rollout, the Ministry of Health Zanzibar played a central role, from recruiting and training CHWs and supervisors, raising awareness in communities, and coordinating supportive supervision for health workers. The standard implementation approach used to scale up the project included community sensitization, recruitment and selection of CHWs and health facility–based supervisors, 10 days of training on the overall program and mobile app, and intensive mentorship for a 3-week period, followed by ongoing monthly supervision and semiannual refresher trainings.

Monitoring, Supervision, and Data-Driven Decision-Making

As a by-product of program implementation, Jamii ni Afya collects extensive data on enrollment, health care usage, and health outcomes. These data are made available on program dashboards that are used by district health management teams to monitor program progress and CHW performance, track health trends, and identify challenges. Supervisors have a mobile app that displays data on the CHWs they supervise, enabling tailored and prompt follow-up and support. We leverage these data in 3 important ways. First, we continuously evaluate the process of program rollout and scale-up to identify successes and areas for improvement [17]. Second, we evaluate the impact of specific components of our program on key outcomes. For example, we found that the number and timing of CHW visits increases the likelihood of health facility delivery [19]. Third, we develop and embed prediction models built using machine learning to tailor our program to individuals and communities [20].

Insights from monitoring and evaluation are used to inform ongoing quality improvement. In 2017, district health managers used the program dashboard to identify that 37% of women were paying for antenatal services, even though they are supposed to be free of charge. The district health team learned that facilities were chronically out of stock for the reagents for tests performed at antenatal visits, and needed to charge clients for services so they could procure these reagents from local pharmacies. With this information, the district health team met with the District Commissioner and presented the data and findings. They were able to obtain a commitment to increase the district health budget to cover the cost of these reagents. Prior to the digital program, district managers knew women were paying for health services but could not back this up with data and were, therefore, never able to advocate for increased funding. Due to this action by the district health team, payments for antenatal care decreased immediately from 37% to 1%.

With Jamii ni Afya established as a fully scaled national government program, there is also increasing interest to leverage the resulting data for other purposes. For example, the RGoZ is exploring how Jamii ni Afya can collect household socioeconomic data during routine visits, which can inform subsidies for the upcoming Universal Health Insurance program.

Sustainability Planning

RGoZ's leadership has been critical to Jamii ni Afya's success and instilled trust in the population for the long-term sustainability of the program. By establishing Jamii ni Afya as part of a national strategy, the government has been able to formally assign oversight and implementation responsibility to government staff so that running the program becomes part of their job description and regular responsibility. The government directly pays the salaries of Jamii ni Afya CHW supervisors (210 individuals), who are health care providers working at health facilities and are 100% dedicated to community health. In addition, 1 person in each of Zanzibar's 11 districts acts as a District Health Promotion Focal Person for Jamii ni Afya. These government employees are responsible for district-level coordination of the program, which has been critical for program oversight, government commitment, and ownership.

The Zanzibar Community Health Strategy Costed Operational Plan developed by the Ministry of Health Zanzibar in 2022 found that the annual operating cost for Jamii ni Afya is TZS 6.4 billion (approximately US $2.73 million), which is equivalent to US $1.70 per capita. Currently, the government is spending US $0.30 per capita on Jamii ni Afya salaries. The remaining costs are covered by donor funds, which funds equipment and working tools, CHW stipends, training, meetings, technology, and hosting costs. The RGoZ has committed to increasing their funding commitments by 25% each year for the next 4 years, until they are fully funding the program by the end of 2026. Of note, the cost per capita is the cost of the entire community health program, which includes the digital health system. We have not isolated the digital system costs since technology is fully integrated in the community health program and cannot be viewed as a stand-alone cost.

Funding and Stakeholder Management

Since the program's inception in 2010, a variety of funders have contributed to the development, piloting, and scale of Jamii ni Afya. From 2010 to 2014, the pilot project was funded by the Bill & Melinda Gates Foundation through their Grand Challenges initiative, which enabled early evidence generation. From 2014 to 2018, the project was funded by the Saving Lives at Birth Grand Challenges Transition to Scale grant, which allowed the program to scale to all districts in Zanzibar and further generate evidence. In 2018, when the program was formally adopted by the RGoZ, funders included Fondation Botnar and the Human Development Innovation Fund. Since 2019, the program also received cofunding from the James Percy Foundation, UNICEF (United Nations Children's Fund), the Conrad N. Hilton Foundation, Google AI for Social Good, the Patrick J. McGovern Foundation, and Jhpiego.

Implementation (Results)

Since August 2021, Jamii ni Afya has been operating on a full national scale. In total, 2300 CHWs and 210 supervisors have been trained to support every household in Zanzibar. As of March 2023, more than 1.5 million people had been registered to the system, representing nearly 80% of Zanzibar's population, and over 320,000 pregnant women and children younger than 5 years have received health visits.

Both the current Jamii ni Afya and prior Safer Deliveries programs focus on education and promotion of health-seeking behavior for maternal and child health and have played a role in positive trends in key health outcomes. The percentage of women delivering at a health facility increased from a baseline of 64% [11] in 2015 to 85% among women registered in the program in 2022, implying a 33% increase. In addition, rates of stunting reduced from 30.4% in 2015 to 17.6% in 2022 [11,21]. While we cannot attribute these reductions to Jamii ni Afya, we are confident that the program's focus on education and promotion of health-seeking behavior for maternal and child health played a role in these significant changes. To directly quantify the Jamii ni Afya's impact on early childhood

development outcomes, our team implemented a nationally representative baseline survey in 2018 with a postimplementation survey planned for late 2023 [22].

There were many factors that led to the success of Jamii ni Afya as a nationally scaled, digital community health program. Briefly, a key element was designing the program by starting with the health challenge and local context rather than the technology. Only when the health challenge was defined and a clear vision for what was possible was conceptualized did we bring in technology, which enabled us to focus on the need, rather than the technology solution. In addition, we leveraged the Principles for Digital Development to guide the technology design process. We also heavily focused on the use of data for decision-making, which amplified the value of the program beyond service delivery, and demonstrated how population-level data could add tremendous value to planning, budgeting, and policy-making at the local, district, and national levels. Finally, extensive collaboration with local and global partners was critical to the success of Jamii ni Afya. The project team engaged with global technology partners, health experts, evaluators, data scientists, and data governance coalitions. Within Zanzibar, collaboration with health experts, universities, political strategists, and communication experts helped to shape the program and support its sustainability. Collaboration with long-term and flexible funders was also key to Jamii ni Afya's success, enabling the project to adapt on the basis of evolving needs and understanding throughout various stages of maturity.

While there were many success factors, there are also a number of challenges and lessons learned. In general, a key lesson is that scaling and sustaining a national digital health program is time-consuming. When we began implementing in 2010, digital health was not a key priority for the RGoZ. It took years to demonstrate the value of the system and get key leaders on board. In addition, it is challenging to successfully scale a digital health intervention when a country has low maturity in foundational aspects of digital health overall (such as digital health leadership and governance, digital health infrastructure, standards and interoperability, and health workforce capacity in digital health). Over the course of the last 13 years, we have been supporting the RGoZ to advance its digital health capacity, which is critical to sustain the Jamii ni Afya program. However, this takes time and significant investment. In addition, managing competing stakeholder interest is a challenge. Often, funders and implementing organizations push for their own interests to be prioritized, which may not be aligned with government priorities. We have been working closely with the Ministry of Health Zanzibar to develop processes and guidelines to manage stakeholder demands in a way that enables continued investment but also ensures that development of the program is aligned with government interest and need. Finally, government capacity and long-term sustainable financing is a challenge, which we are addressing through a formal transition plan, and describe further in the *Discussion* section.

The current focus on Jamii ni Afya is full transition to the RGoZ. Working with partners, the government has developed a transition plan that outlines the financial, technological, and operational aspects of the program that will be fully absorbed by the government by 2026. D-tree International and partners are supporting skills transfer and systems development to enable this transition, recognizing that financing and management of a nationally scaled digitally enabled community health program is a major undertaking, and strong accompaniment from nongovernmental organizations is critical to the effective handover of such a system.

Discussion

This implementation report outlines the evolution of Zanzibar's national, digitally enabled community health program. Since 2010, the program grew from a small pilot to a nationally scaled, government-owned initiative that is set to be fully institutionalized within the national health system. During this time, we have learned many lessons and best practices that may be relevant to governments and practitioners designing, scaling, and institutionalizing digitally enabled community health programs.

Community health programs have traditionally been driven by vertical health programs (ie, HIV/AIDS, family planning, or maternal health) and focus on specific health service delivery [9,23]. With Jamii ni Afya, we focused on designing a strong community health system, independent of the types of services being delivered, which serves as a strong foundation on which to add service delivery areas. The Zanzibar Community Health Strategy is a critical document that not only formalizes Jamii ni Afya as a government initiative, but also outlines critical systems-level criteria to guide the program. This includes defining the supervision structure, selection criteria, and training and stipend payments, which is aligned with the World Health Organization's guidelines for community health [10].

Just as it is important to invest in building strong foundations in the community health system, it is critical to design digital health initiatives that are integrated and harmonized with the broader digital health ecosystem. A recent article by Karamagi et al [24] describes the alarming lack of coordination, integration, scalability, and sustainability of digital health interventions in sub-Saharan Africa. Through the design of Jamii ni Afya, we were intentional in ensuring that the digital system was interoperable with other digital health systems such as DHIS2 in order to enable the transfer of data between systems. The Jamii ni Afya team was also closely involved in the development of the Zanzibar Digital Health Strategy, which features Jamii ni Afya as a core digital health intervention. Jamii ni Afya was among the first systems to adopt the Zanzibar Health Interoperability Layer, which lays the foundation for seamless data exchange across systems. These intentional design decisions, coupled with integrating the program into government policy, have set the stage to ensure long-term coordination and synergy with the evolving digital health landscape in Zanzibar.

Having the right leadership is recognized as a critical lever to accelerate the institutionalization of a product or service [25]. Strong leadership, ranging from the highest levels of government to local community leaders, may be the single-most important success factor for Jamii ni Afya. Extensive work was carried out early on to establish buy-in and trust at the community level. Community sensitization was built into the national rollout plan, leveraging shehas (ward leaders) to champion the initiative

within their communities. Sensitization activities were led by district- and national-level government staff who introduced the program as a government-led initiative, and CHWs have official government-issued identification cards that provide increased credibility. These initiatives were critical in building public trust and ownership of Jamii ni Afya from the beginning, resulting in 90% of the population being registered in the program. We also focused on high-level political support to champion this initiative and set the stage for full government ownership and sustainable financing. In February 2023, the Minister of Health Zanzibar presented Jamii ni Afya to the Zanzibar President who pledged his commitment to champion the institutionalization of this initiative. While this support is encouraging, one lesson learned is the importance of high-level political engagement early on. The Jamii ni Afya team initially focused on community, district, and departmental support within the Ministry of Health Zanzibar in order to build buy-in and support, and only after the program was established, began lobbying higher levels of government for their support. Earlier engagement with these higher levels may have expedited the institutionalization process.

The evolution of Jamii ni Afya has lasted nearly 13 years and is now in a strong position to be institutionalized within Zanzibar's health system. Each stage of the project—beginning with a small pilot to demonstrate the potential of this model, gaining consensus for government adoption, and integrating the program into government policies—was critical to build demand and formalize Jamii ni Afya as a government initiative. Often, donors expect that health systems strengthening initiative, including digital health, should be conceptualized, piloted, scaled, and institutionalized within a short period of a few years. However, our experience is consistent with that of other studies and shows that significant digital health systems transformations take time and require funding and partnerships to accompany governments on their journey to scale [26].

One of the most significant challenges for Jamii ni Afya's institutionalization is sustainable financing, which is a common challenge for community health programs across many low- and middle-income countries [27]. The RGoZ has committed to full institutionalization of Jamii ni Afya by 2026, which is an ambitious target and will require substantial political commitment and innovative solutions. The Jamii ni Afya program's team is currently working on a number of strategies to support the government to fully finance the program within 4 years. One way is to increase demand for the program outside of the health sector. Jamii ni Afya champions are in discussion with the Ministry of Finance Zanzibar, Ministry of Social Welfare, Ministry of Agriculture, and Ministry of Education to discuss how existing or new data from Jamii ni Afya could be valuable to them, or the potential of expanding the package of services delivered by CHWs. By extending the value of the program outside of the Ministry of Health Zanzibar, additional government entities can contribute toward operating costs, thus reducing the burden on any one particular ministry and reducing duplication of efforts. We are also in the process of integrating Jamii ni Afya into the government's upcoming Universal Health Insurance scheme, which will be rolled out in the next 1-2 years. The Jamii ni Afya digital platform can be leveraged to collect data on household economic status and feedback on health facility quality, and CHWs can support registration of households into the program. In turn, the community health program can be partially financed by revenue generated through the Universal Health Insurance scheme. It has been reported that many national health insurance schemes fail to adequately engage the informal sector [28]. Leveraging CHWs to support these efforts in Zanzibar could both increase participation in the informal sector and support operational costs for the community health program. There is also increasing support from multilateral donors who are increasingly investing in digitally enabled community health systems and provide funding directly to governments to decrease financial reliance on nongovernmental partners. This multipronged financing approach is aligned with a recent review that found that an adaptive mix of health financing mechanisms is necessary for low- and middle-income countries [29].

In conclusion, Jamii ni Afya represents one of the world's first examples of a nationally scaled digitally enabled community health program. This implementation report outlines the evolution of this program, from the pilot to the national scale and institutionalization. While Zanzibar has a relatively small population and is geographically isolated, the challenges they face are similar to many other settings. Indeed, based on the experience in Zanzibar, D-tree International is replicating this approach by supporting government-led digital community health initiatives in Mainland Tanzania and Zambia. However, drawing on work from Greenhalgh et al [30], we also recognize that the spread and scale-up of innovations is a combination of technical, ecological, and social aspects and require social science approaches to understand and facilitate change based on local contextual factors and incentives [30]. The authors hope that the experiences and lessons learned described in this implementation report will be helpful to others working to implement digitally enabled community health systems at scale. In addition, given the maturity of Jamii ni Afya, the relatively small size of Zanzibar's population, and the significant health need, there is an opportunity for Zanzibar to become a digital health implementation research hub in order to test new and promising innovations, research various implementation models, and develop best practices that can be applicable in other settings.

Authors' Contributions

EL made substantial contributions to the conception and design, drafted the manuscript, revised it critically, and granted final approval for the version submitted. SS, IM, and AWAM revised the manuscript critically and granted final approval for the version submitted. GVRB made substantial contributions to the conception and design, revised the manuscript critically, and granted final approval of the version submitted. MM revised the manuscript critically and granted final approval for the version

submitted. IF made substantial contributions to the conception and design, revised the manuscript critically, and granted final approval for the version submitted. HH made substantial contributions to the conception and design, revised the manuscript critically, and granted final approval for the version submitted. RL made substantial contributions to the conception and design, revised the manuscript critically, and granted final approval of the version submitted.

Conflicts of Interest

None declared.

References

1. Quality in Primary Health Care. World Health Organization. 2018. URL: https://www.who.int/docs/default-source/primary-health-care-conference/quality.pdf [accessed 2023-08-30]
2. Kruk M, Gage A, Joseph N, Danaei G, García-Saisó S, Salomon J. Mortality due to low-quality health systems in the universal health coverage era: a systematic analysis of amenable deaths in 137 countries. The Lancet 2018 Nov;392(10160):2203-2212 [FREE Full text] [doi: 10.1016/s0140-6736(18)31668-4]
3. Global strategy on human resources for health: Workforce 2030. World Health Organization. 2016. URL: https://apps.who.int/iris/bitstream/handle/10665/250368/9789241511131-eng.pdf [accessed 2023-08-30]
4. World Health Organization. WHO guideline on health policy and system support to optimize community health worker programmes. Geneva: World Health Organization; 2018.
5. Cometto G, Ford N, Pfaffman-Zambruni J, Akl E, Lehmann U, McPake B, et al. Health policy and system support to optimise community health worker programmes: an abridged WHO guideline. Lancet Glob Health 2018 Dec;6(12):e1397-e1404 [FREE Full text] [doi: 10.1016/s2214-109x(18)30482-0]
6. World Health Assembly. Community health workers delivering primary health care: opportunities and challenges. World Health Organization. 2019. URL: http://apps.who.int/gb/ebwha/pdf_files/WHA72/A72_R3-en.pdf [accessed 2023-02-27]
7. Scott K, Beckham S, Gross M, Pariyo G, Rao KD, Cometto G, et al. What do we know about community-based health worker programs? A systematic review of existing reviews on community health workers. Hum Resour Health 2018 Aug 16;16(1):39 [FREE Full text] [doi: 10.1186/s12960-018-0304-x] [Medline: 30115074]
8. Dahn B, Woldemariam A, Perry H, Maeda A, von Glahn D, Panjabi R, et al. Strengthening Primary Health Care through Community Health Workers: Investment Case and Financing Recommendations. CHW Central. 2015. URL: https://chwcentral.org/resources/strengthening-primary-health-care-through-community-health-workers-investment-case-and-financing-recommendations/ [accessed 2023-02-27]
9. Hodgins S, Kok M, Musoke D, Lewin S, Crigler L, LeBan K, et al. Community health workers at the dawn of a new era: 1. Introduction: tensions confronting large-scale CHW programmes. Health Res Policy Syst 2021 Oct 12;19(Suppl 3):109 [FREE Full text] [doi: 10.1186/s12961-021-00752-8] [Medline: 34641886]
10. Recommendations on digital interventions for health system strengthening. World Health Organization. 2019. URL: https://www.who.int/publications/i/item/9789241550505 [accessed 2023-08-30]
11. Ministry of Health, Community Development, Gender, Elderly and Children Dar es Salaam, Ministry of Health Zanzibar, National Bureau of Statistics Dar es Salaam, Office of Chief Government Statistician Zanzibar, ICF Rockville, Maryland USA. Tanzania Demographic and Health Survey and Malaria Indicator Survey 2015-2016. 2016. URL: https://dhsprogram.com/pubs/pdf/fr321/fr321.pdf [accessed 2023-08-30]
12. Ministry of Health, Community Development, Gender, Elderly and Children (MoHCDGEC) [Tanzania Mainland], Ministry of Health (MoH) [Zanzibar], Tanzania Food and Nutrition Centre (TFNC), National Bureau of Statistics (NBS), Office of the Chief Government Statistician (OCGS) [Zanzibar], UNICEF. Tanzania National Nutrition Survey using SMART Methodology (TNNS) 2018. 2018. URL: https://www.unicef.org/tanzania/media/2141/file/Tanzania%20National%20Nutrition%20Survey%202018.pdf [accessed 2023-08-30]
13. NCD Survey Report Main findings from the National Non-Communicable Disease Risk Factor Survey 2011. Ministry of Health Zanzibar. 2011. URL: https://cdn.who.int/media/docs/default-source/ncds/ncd-surveillance/data-reporting/zanzibar/steps/2011_zanzibar_steps_report.pdf?sfvrsn=368ed243_3&download=true [accessed 2023-08-30]
14. Revolutionary GOZ. Ministry of Health, Health Sector Strategic Plan IV (2020/21 - 2024/25). Ministry of Health, Republic of Uganda. URL: https://mohz.go.tz/eng/ministry-different-strategies/ [accessed 2023-06-01]
15. Perrin Franck C, Babington-Ashaye A, Dietrich D, Bediang G, Veltsos P, Gupta PP, et al. iCHECK-DH: guidelines and checklist for the reporting on digital health implementations. J Med Internet Res 2023 May 10;25:e46694 [FREE Full text] [doi: 10.2196/46694] [Medline: 37163336]
16. Battle J, Farrow L, Tibaijuka J, Mitchell M. mHealth for Safer Deliveries: A mixed methods evaluation of the effect of an integrated mobile health intervention on maternal care utilization. Healthc (Amst) 2015 Dec;3(4):180-184 [FREE Full text] [doi: 10.1016/j.hjdsi.2015.10.011] [Medline: 26699340]
17. Fulcher I, Nelson A, Tibaijuka J, Seif SS, Lilienfeld S, Abdalla OA, et al. Improving health facility delivery rates in Zanzibar, Tanzania through a large-scale digital community health volunteer programme: a process evaluation. Health Policy Plan 2021 Mar 16;35(10):1-11 [FREE Full text] [doi: 10.1093/heapol/czaa068] [Medline: 33263749]

18. Principles for Digital Development. URL: https://digitalprinciples.org/ [accessed 2023-02-27]
19. Hentschel E, Russell A, Said S, Tibaijuka J, Hedt-Gauthier B, Fulcher I. Identifying programmatic factors that increase likelihood of health facility delivery: results from a community health worker program in Zanzibar. Matern Child Health J 2022 Sep;26(9):1840-1853 [FREE Full text] [doi: 10.1007/s10995-022-03432-3] [Medline: 35386028]
20. Fredriksson A, Fulcher I, Russell A, Li T, Tsai YT, Seif SS, et al. Machine learning for maternal health: predicting delivery location in a community health worker program in Zanzibar. Front Digit Health 2022;4:855236 [FREE Full text] [doi: 10.3389/fdgth.2022.855236] [Medline: 36060544]
21. Ministry of Health Dodoma, Ministry of Health Zanzibar, National Bureau of Statistics Dodoma, Office of Chief Government Statistician Zanzibar, The DHS Program ICF Rockville, Maryland, USA. Demographic and Health Survey and Malaria Indicator Survey 2022. Key Indicators Report. 2023. URL: https://dhsprogram.com/methodology/survey/survey-display-578.cfm [accessed 2023-02-27]
22. Russell A, Hentschel E, Fulcher I, Ravà MS, Abdulkarim G, Abdalla O, et al. Caregiver parenting practices, dietary diversity knowledge, and association with early childhood development outcomes among children aged 18-29 months in Zanzibar, Tanzania: a cross-sectional survey. BMC Public Health 2022 Apr 15;22(1):762 [FREE Full text] [doi: 10.1186/s12889-022-13009-y] [Medline: 35428252]
23. Lu C, Palazuelos D, Luan Y, Sachs S, Mitnick C, Rhatigan J, et al. Development assistance for community health workers in 114 low- and middle-income countries, 2007–2017. Bull World Health Organ 2019 Nov 01;98(1):30-39 [FREE Full text] [doi: 10.2471/blt.19.235499]
24. Karamagi H, Muneene D, Droti B, Jepchumba V, Okeibunor JC, Nabyonga J, et al. eHealth or e-chaos: the use of digital health interventions for health systems strengthening in sub-Saharan Africa over the last 10 years: a scoping review. J Glob Health 2022 Dec 03;12:04090 [FREE Full text] [doi: 10.7189/jogh.12.04090] [Medline: 36462201]
25. Wilson K, Gertz B, Arenth B, Salisbury N. The journey to scale: Moving together past digital health pilots. Seattle, WA: PATH; 2014.
26. Benjamin K, Potts HW. Digital transformation in government: lessons for digital health? Digit Health 2018;4:2055207618759168 [FREE Full text] [doi: 10.1177/2055207618759168] [Medline: 29942624]
27. Masis L, Gichaga A, Zerayacob T, Lu C, Perry H. Community health workers at the dawn of a new era: 4. Programme financing. Health Res Policy Syst 2021 Oct 12;19(Suppl 3):107 [FREE Full text] [doi: 10.1186/s12961-021-00751-9] [Medline: 34641893]
28. Fenny A, Yates R, Thompson R. Strategies for financing social health insurance schemes for providing universal health care: a comparative analysis of five countries. Glob Health Action 2021 Jan 01;14(1):1868054 [FREE Full text] [doi: 10.1080/16549716.2020.1868054] [Medline: 33472557]
29. Domapielle M, Sumankuuro J, Bebelleh F. Revisiting the debate on health financing in Low and Middle-income countries: an integrative review of selected models. Int J Health Plann Manage 2022 Nov;37(6):3061-3074 [FREE Full text] [doi: 10.1002/hpm.3566] [Medline: 36030531]
30. Greenhalgh T, Papoutsi C. Spreading and scaling up innovation and improvement. BMJ 2019 May 10;365:l2068 [FREE Full text] [doi: 10.1136/bmj.l2068] [Medline: 31076440]

Abbreviations

CHT: Community Health Toolkit
CHW: community health worker
DHIS2: District Health Information System
iCHECK-DH: Guidelines and Checklist for the Reporting on Digital Health Implementations
NGO: nongovernmental organization
RGoZ: Revolutionary Government of Zanzibar
UHC: universal health coverage
UNICEF: United Nations Children's Fund

Edited by C Lovis, C Perrin; submitted 13.04.23; peer-reviewed by S Delaigue, N Ash, Y Sapanel, M Randriambelonoro; comments to author 19.05.23; revised version received 19.06.23; accepted 18.08.23; published 09.10.23.

Please cite as:
Layer E, Slim S, Mussa I, Al-Mafazy AW, Besana GVR, Msellem M, Fulcher I, Hornung H, Lampariello R
The Journey of Zanzibar's Digitally Enabled Community Health Program to National Scale: Implementation Report
JMIR Med Inform 2023;11:e48097
URL: https://medinform.jmir.org/2023/1/e48097
doi:10.2196/48097
PMID:37812488

Implementation Report

Implementing Clinical Information Systems in Sub-Saharan Africa: Report and Lessons Learned From the MatLook Project in Cameroon

Georges Bediang[1], Prof Dr Med

Faculty of Medicine and Biomedical Sciences, Université de Yaoundé, Yaoundé, Cameroon

Corresponding Author:
Georges Bediang, Prof Dr Med
Faculty of Medicine and Biomedical Sciences
Université de Yaoundé
PO Box 1364
Yaoundé
Cameroon
Phone: 237 699588574
Email: bediang@yahoo.com

Abstract

Background: Yaoundé Central Hospital (YCH), located in the capital of Cameroon, is one of the leading referral hospitals in Cameroon. The hospital has several departments, including the Department of Gynecology-Obstetrics (hereinafter referred to as "the Maternity"). This clinical department has faced numerous problems with clinical information management, including the lack of high-quality and reliable clinical information, lack of access to this information, and poor use of this information.

Objective: We aim to improve the management of clinical information generated at the Maternity at YCH and to describe the challenges, success factors, and lessons learned during its implementation and use.

Methods: Based on an open-source hospital information system (HIS), this intervention implemented a clinical information system (CIS) at the Maternity at YCH and was carried out using the HERMES model—the first part aimed to cover outpatient consultations, billing, and cash management of the Maternity. Geneva University Hospitals supported this project, and several outcomes were measured at the end. The following outcomes were assessed: project management, technical and organizational aspects, leadership, change management, user training, and system use.

Implementation (Results): The first part of the project was completed, and the CIS was deployed in the Maternity at YCH. The main technical activities were adapting the open-source HIS to manage outpatient consultations and develop integrated billing and cash management software. In addition to technical aspects, we implemented several other activities. They consisted of the implementation of appropriate project governance or management, improvement of the organizational processes at the Maternity, promotion of the local digital health leadership and performance of change management, and implementation of the training and support of users. Despite barriers encountered during the project, the 6-month evaluation showed that the CIS was effectively used during the first 6 months.

Conclusions: Implementation of the HIS or CIS is feasible in a resource-limited setting such as Cameroon. The CIS was implemented based on good practices at the Maternity at YCH. This project had successes but also many challenges. Beyond project management and technical and financial aspects, the other main problems of implementing health information systems or HISs in Africa lie in digital health leadership, governance, and change management. This digital health leadership, governance, and change management should prioritize data as a tool for improving productivity and managing health institutions, and promote a data culture among health professionals to support a change in mindset and the acquisition of information management skills. Moreover, in countries with a highly centralized political system like ours, a high-level strategic and political anchor for such projects is often necessary to guarantee their success.

(JMIR Med Inform 2023;11:e48256) doi:10.2196/48256

KEYWORDS

implementation report; challenges; success factors; Sub-Saharan Africa; Cameroon; healthcare; health care; clinical information; information management; clinical information systems; hospital information systems; data governance

Introduction

Cameroon is a transitional country located in Central Africa. As a resource-limited country, its health system faces many challenges and barriers [1,2]; these include inadequate quality of care, human resources, health infrastructure, poor drug supply, poor health financing, and poor health information management. Yaoundé Central Hospital (YCH) is located in Cameroon's capital city; it is one of the country's leading national referral hospitals. YCH has several departments, including the Department of Gynecology-Obstetrics, also known as (and hereinafter referred to as) "the Maternity." The management of clinical information at the Maternity has multiple limitations including those identified by the World Health Organization's classification of digital health interventions: lack of high-quality and reliable clinical information (issues with data completeness, integrity, and confidentiality), lack of access to the information, and its insufficient use for management and decision-making [3]. This poorly managed clinical information prevents clinicians from accessing information regarding diagnoses, treatments, prescriptions, and previous care [4]. This results in fragmented processes, reduced quality and safety of care, and increased cost of care. Implementing and strengthening robust hospital information systems (HISs) or clinical information systems (CISs) is one of the solutions to these problems [5-8]. The MatLook Project was initiated between the Geneva University Hospitals (HUG) and the YCH to implement an electronic CIS to strengthen the HIS in the Maternity at YCH. This implementation report follows the Guidelines and Checklist for the Reporting on Digital Health Implementations (iCHECK-DH) [9].

Methods

Objectives

The general objective of the MatLook Project was to improve the management of clinical information generated at the Maternity at YCH. This implementation report describes the results of the implementation and the associated challenges, success factors, and lessons learned. The primary outcome of this implementation was the completion of the parameterization of MediBoard [10] and services developed, and determination of the level of the deployment of the solution. Secondary outcomes were the development of a procedure manual and determination of the level of its implementation, implementation of training sessions for users, determination of the level of implementation of change management support for users, determination of the level of use of the solution, establishment of project governance or management, development of local leadership to promote the use of the solution, and use of this information for the management of the Maternity or the hospital.

Technical Design

The objectives of this project were aligned with YCH's strategy, one of its goals being to use ICT to improve YCH's productivity. This project corresponds to categories H and K (electronic medical record and facility management information system, respectively) of World Health Organization's classification of digital health interventions [3]. At the time this project was implemented, Cameroon did not have a national digital health strategy.

The project was conducted using the HERMES method [11]. This project management method is used in IT, service or product development, and organizational adaptation. It consists of 4 phases: initialization, design, implementation, and deployment. Each step comprises tasks or activities to be performed by roles to achieve the expected results. Each stage leads to a decision or milestone that enables the project to proceed to the next phase. In addition to these phases, plans were developed for risk management, marketing and communication (web portals, newsletters, and information sessions), project quality assurance, and solution testing.

The project was organized into 2 parts. The first part aimed to cover outpatient consultations, billing, and maternity cash management. Billing and cash management were to be handled through the additional development of billing and cash management software that interoperates with this CIS. This first part, including all the 4 phases of the HERMES model, cross-functional project management activities, and users' training, lasted approximately 15 months. The second part aimed to cover other care processes, such as hospitalization management, delivery management, surgery management, etc.

From a technical point of view, the project consisted in adapting and parameterizing an existing open-source hospital information system called MediBoard [10]. The choice of this solution was based on a comparative study of several HISs with the advantage of being open-source with a dynamic community, the existence of a French version, the fact that several functionalities cover the needs of a hospital located in a resource-limited setting such as Cameroon, and the possibility of making the necessary adaptations relatively quickly [12]. It was also based on the experience of a successful implementation in Mali [6]. MediBoard is an open-source HIS developed in a modular and multilayer web architecture, using Apache, PHP, MySQL, XML, XHTML, JavaScript, CSS, Prototype.js, Smarty, and PEAR. It contains several functionalities such as management of administrative, medical, and nursing patient files; prescription management; consultation appointment management; hospitalization management; billing and accounting management (this service was not adapted to the processes in place, where, for example, people pay before they receive treatment); stock management; human resources management; meal management; quality assurance; incident management; etc.

Target

The Maternity at YCH was chosen as the pilot unit to implement the solution. It is the second largest maternity unit in Yaoundé after the Yaoundé Gynaecology, Obstetrics and Pediatrics Hospital, which specializes in mother and child care (detailed data not available). This clinical department registers approximately 20,000 patients per year (65% of which are registered through outpatient consultations), 3500 deliveries, and 1000 surgical interventions (63% of which are caesarean deliveries, the remaining 37% of them being ectopic pregnancies, uterine revisions, hysterectomies, myomectomies, cystectomies, etc).

The targets of this intervention were the staff at the Maternity at YCH, including the managers and administrators, health professionals, the team responsible for billing and cash management, and the patients.

Data

Cameroon has a law (n° 2010/012, December 21, 2010) on cybersecurity and cybercrime, which protects and ensures respect for the privacy of individuals and punishes offences related to the use of information and communication technologies in Cameroon.

Data were collected using computers in consultation rooms, doctors' offices, and treatment rooms. The data were processed and stored at the YCH (data owner) on a data server acquired for this project. The data were disseminated through the intranet network implemented for this project. Several security measures were implemented to ensure data security: physical security (secure rooms dedicated to the server and switches) and computer security and confidentiality (access to the wired intranet network via a dynamic host configuration protocol, implementation of firewalls, management of user profiles, and control of access rights—authentication and authorization—via logins and passwords). Based on their level of confidentiality, the people who had access to these data were reception and triage nurses, doctors, nurses, administrative and financial staff, and patients.

Interoperability

MediBoard integrates technical interoperability standards such as HL7 (Health Level 7; technical specifications for computerized exchange of clinical, financial, and administrative data between HISs) and HPRIM (harmoniser et promouvoir l'informatique médicale [English: harmonizing and promoting medical informatics] for transmitting information regarding biological examinations in France). It also integrates semantic interoperability standards such as *ICD-10* (*10th revision of the International Classification of Diseases*) and CCAM (Classification Commune des Actes Médicaux [English: Common Classification of Medical Procedures]), which lists all medical procedures to be performed by doctors, midwives, and dental surgeons.

Participating Entities

At the national level, this project was supported exclusively by the direction of the YCH through the signing of an authorization by the director to carry out a feasibility study and the signing of a decision authorizing the implementation of the project in this hospital. This local institutional support, the needs and requests formulated by the beneficiaries, the potential of the YCH in terms of equipment and human resources, the existence of local expertise in the field of digital health, the individuals in charge of operational management of the project, and the financial guarantees enabled us to implement the project. The project was implemented within the framework of existing cooperation agreements between the YCH, which committed to financing one-third of the budget, and the HUG, which committed to funding two-thirds of the budget. The government was not directly involved because the framework was the existing cooperation agreement between these 2 hospitals.

Budget Planning

This project was estimated at CHF 60,000 (US $65,144.64): CHF 8000 (US $8685.95) for the adaptation of the MediBoard software; CHF 33,000 (US $35,829.55) for investments (intranet network, servers, computers, switches, etc), communication, and marketing of the project; CHF 14,000 (US $15,200.42) for operations (project management, change management, management of steering, and working meetings with the Maternity user group); CHF 2000 (US $2171.49) for training and support of users; and finally CHF 3000 (US $3257.23) for the evaluation of the solution by an external evaluator of the project. This budget was intended to cover the first and second parts of the project.

Sustainability

To ensure the sustainability of this project, several actions have been taken, including the alignment of the project's objectives with those of YCH and its implementation with the agreement of the director of YCH; the commitment of the director to fund one-third of the budget; the establishment of a steering committee including all YCH stakeholders; the integration of the YCH IT manager as the deputy project manager; the creation of a user group to assist the project group in the design and the implementation of the project; the designation of local champions to promote the use of the CIS; and finally, the official launch (supported by a document signed by the director) of this CIS at the Maternity.

Ethical Considerations

The study does not require ethical approval since it focuses on the implementation of a clinical information system based on an open-source hospital information system. No individual patient data was collected as part of this project.

Implementation (Results)

Coverage

The first part of the project covered the Maternity's outpatient consultations and billing and cash management. The processes covered were the administrative and financial management of patients (administrative data, billing, and payments); management of appointments and clinical information (clinical observations, requests for laboratory and radiological tests, management of the results of these tests, prescriptions for medical consumables, drugs, and medical procedures). The second part did not take place due to nonpayment of one-third of the funding from the YCH.

Outcomes

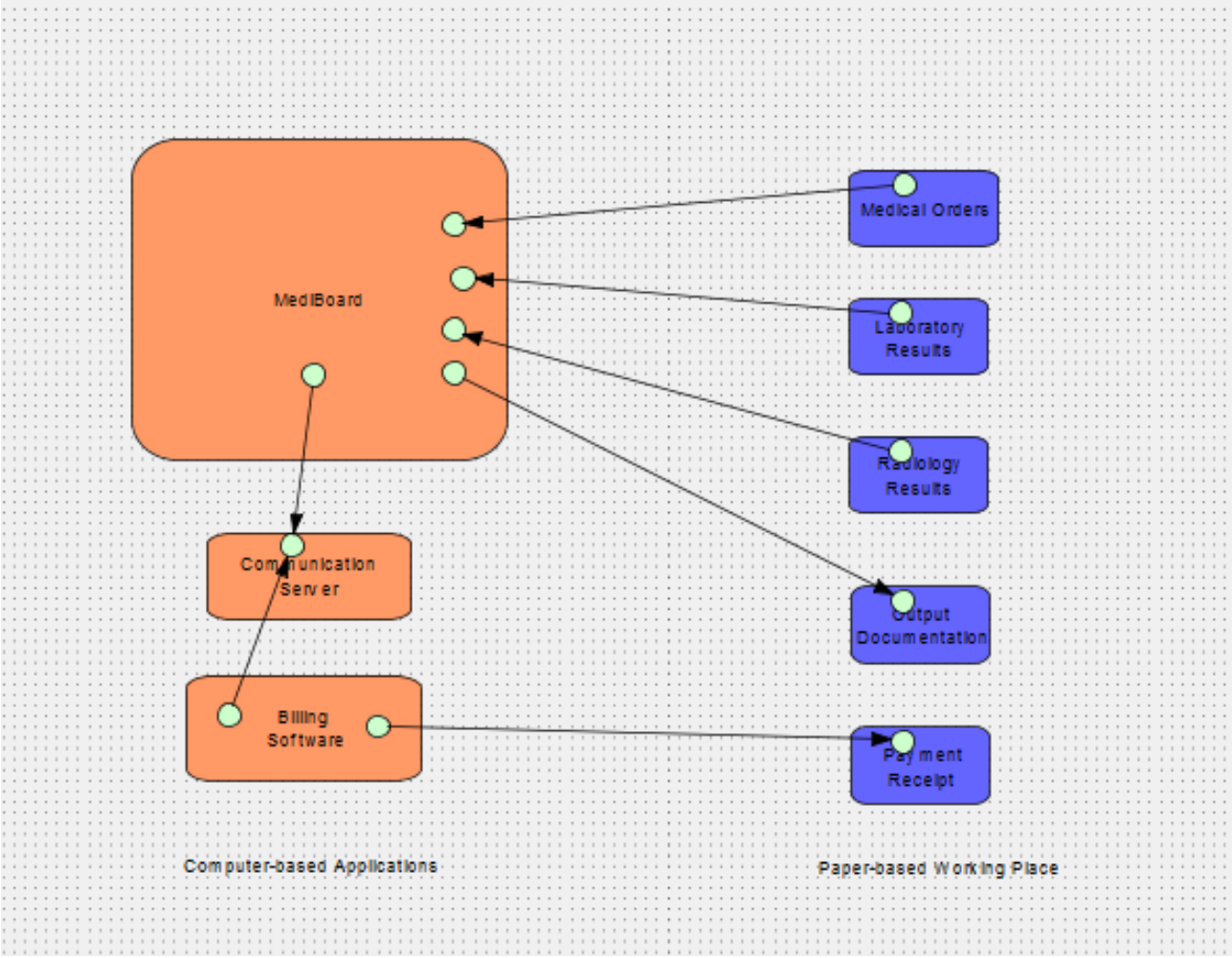
The MediBoard software was adapted to manage outpatient consultations. The software parameterization activities included filling in the YCH metadata; adapting forms for entering administrative and sociodemographic data of patients, clinical observations, laboratory and radiology test requests, collection of test result data, medical consumables, and prescription for drugs and medical procedure; defining user profiles (doctor-head of department, gynecologist, resident, reception nurse, student, and secretary) and creating user accounts for maternity health care professionals.

A complementary billing and cash management module, integrating local coding of medical procedures was developed, validated ad hoc, and integrated into the MediBoard HIS (Figure 1) [13]. This software had several functionalities, namely invoice management (creation, modification, and possibility to pay all or part of the invoice); invoice quote management (creation, modification, and transformation of invoice quotes); medical procedure management (editing of medical procedures carried out or the tariffs of these medical procedures); and cash and accounting management (statements of financial entries by the cashier by medical procedure, service, practitioner, period, etc). These services have been deployed at the Maternity at YCH.

An intranet network with a capacity of 40 computers was implemented, and equipment (1 server and 20 laptops) was provided. Security was ensured by physical measures (restriction of access to servers, switches, and consultation rooms) and information and communication technology (ICT) measures (restriction of access to the local network and access to the user account via logins and passwords).

The solution was deployed to cover the Maternity's outpatient consultations and billing and cash management and was used over 6 months.

Figure 1. The architecture of the clinical information system of the Department of Gynecology-Obstetrics at Yaoundé Central Hospital (logical level based on the 3LGM^2 model).



To ensure proper functioning of the solution, a user guide and a procedure manual specifying required organizational architecture, processes covered, rules, roles and privileges, and options for entering medical data during the consultation were developed. This procedure manual made it possible, for example, to redefine the workflow for the patient coming for an outpatient consultation.

Users (the staff involved) were regularly trained (before the start of the use of the system and at each staff turnover; this was due to high staff turnover) on the new organizational processes and for proper use of the CIS. This training was provided by a dedicated staff explicitly recruited in the project for this task.

In total, 80 people were trained, including senior gynecologists, gynecology residents, medical students, nurses in charge of reception and recording patients' parameters, and administrative and financial staff. This was a 100% increase on the initial target of people who could benefit from the training during this stage of the project (part 1).

Support staff was recruited to provide specific support to users of the CIS. Computer maintenance was shared between the project team and the YCH IT staff.

An evaluation of the use of this system carried out 6 months after the use of the CIS revealed 113 users, 1278 patient records,

and 960 medical consultations recorded, corresponding to a financial income of approximately CHF 8000 (US $8685).

In terms of project governance or management, the project was organized as follows: a steering committee comprising the hospital's administrative, business, and financial managers to steer the project; a project team comprising external and internal skills in medical informatics, informatics, and project management to implement the project; and a user group comprising health professionals at the Maternity (1 department head physician; 2 gynecologists; 1 anesthesiologist; 4 care coordinators for the outpatient consultations, delivery room, hospitalization, and operating room; 1 nurse; and 1 midwife) to monitor the project activities and to verify that they meet the users' needs. A marketing and risk management plan for the project has been drawn up. A communication based on the distribution of posters presenting the advantages of the new solution was carried out for the health care professionals and the patients.

Regarding leadership, the 2 heads of department for maternity units A and B were involved as local champions (health professionals of the Maternity who have adopted IT and are ready to provide leadership for their integration in health care in motherhood) to promote the use of the CIS.

Due to the short evaluation period, we could not objectify using the dashboard indicators in the management of the Maternity or the hospital.

Lessons Learned

Success Factors

Several factors contributed positively to the progress of the project—these include the feasibility and implementation agreements of the MatLook project issued by the YCH management; the implementation of a project organization integrating the steering committee, the project group, and the user group; the management of the project based on a robust methodology (HERMES model); the involvement of local champions to ensure leadership and to promote the use of this system; the success of software parameterization and development activities; the realization of several communication and marketing activities for the project and the ongoing training and support of users; and the feasibility and implementation agreements of the MatLook project issued by the YCH management.

Challenges

Throughout the project, several challenges or barriers were encountered. These included a change of director of the YCH (3 months after the official launch of the project); lack of leadership from local champions; lack of standard care procedures or protocols; resistance to change; insufficient ICT skills among maternity staff; lack of staff to enter data into the electronic CIS; high staff turnover leading to an increase in the frequency of training sessions on the use of the CIS; poor promotion and valorization of this project within the institution; and an increase in workload due to the diversity of supports to be completed (paper-based service registers coupled with the electronic CIS).

Finally, another challenge was about the reliability and confidentiality of the data collected and the level of responsibility of residents and students for these data. Due to a lack of ICT skills, senior gynecologists gave their logins and passwords to residents or students to document patients' case files in the electronic CIS on their behalf. This last problem, relating to the exchange of logins and passwords among gynecologists, residents, and students, was identified as soon as the CIS was introduced. To mitigate this, we carried out an awareness-raising campaign among users, demonstrating the risks of this practice and urging them to stop. In addition, we held a meeting with the user group to find more sustainable solutions. The solutions proposed were organizing workshops to strengthen gynecologists' data entry skills; making consultation notebooks with tracers so that stubs could be used for data entry after patients' encounters; and recruiting medical secretaries dedicated to this data entry activity. Unfortunately, none of these measures could be implemented due to a lack of resources.

Budget Report

Of the expected CHF 60,000 (US $65,144.64), the project received CHF 40,000 (US $43,360.52; two-thirds of the expected funding) from HUG. These funds were allocated as follows: CHF 8000 (US $8685.95) for the adaptation of the MediBoard software (part 1), CHF 18,500 (US $20,054.24) for the investments (intranet network, computer equipment, and communication and marketing of the project), CHF 9000 (US $9756.12) for operations (project management, change management, management of steering and working meetings, and incentives for health professionals to document cases), CHF 3000 (US $3257.23) for training and support of users, and finally, CHF 1500 (US $1626.02) for the transport of some equipment (servers and computers) from Geneva to Yaoundé and bank account fees. The CHF 20,000 (US $21,680.26; one-third) that was to be paid by the YCH was not paid. This money was to be used as follows: CHF 4000 (US $4336.05) for the adaptation of the MediBoard software (part 2), CHF 4000 (US $4336.05) for the investments (computer equipment, communication, and marketing of the project), CHF 6000 (US $6504.08) for operations (project management, change management, management of steering and working meetings, and incentives for health professionals to document cases), CHF 3000 (US $3252.04) for training and support of users, and finally, CHF 3000 (US $3252.04) for the evaluation of the solution.

Unintended Consequences

During this project and after implementing some approaches for change management (setting up a user group, identifying local champions, training users, implementing new processes based on the procedure manual, and setting up user support), we experienced an unintended or unexpected event—staff reluctance to use the system after implementation. The reason was the lack of financial motivation. This situation required the project management team to implement a special documentation gratification (per diem introduced to encourage staff to enter patient information into the system). At the end of this documentation gratification fund (approximately 6 months after

the beginning of the use of the CIS), users stopped entering data. In addition, due to the lack of funding from the YCH, part 2 of the project could not be continued, which raises the issue of trust and institutional continuity in sub-Saharan Africa.

Discussion

Implementing HIS or CIS is feasible in a resource-limited country such as Cameroon. The management of this process is complex and critical to the success of the project.

The implementation of this CIS at the Maternity at YCH was carried out following the defined and approved institutional framework (feasibility and implementation agreements issued by the YCH management, mobilization, and cofunding of the HUG partner) and by relying on a methodology (HERMES phase model) adapted to the implementation of such projects. This project enabled the deployment of a functional CIS at the Maternity at YCH covering (in part 1) outpatient consultations and billing and cash management.

As with any project of this type in our setting, where the implementation of HIS or CIS is still at the embryonic stage, this project registered successes but also many challenges, which have been described in detail in the *Implementation (Results)* section.

Publications describing such implementations in Africa have highlighted similar difficulties: low staff involvement, lack of project management skills, insufficient ICT skills among users, lack of political and administrative support for the project, lack of an interoperability framework, low funding, and lack of logistical support and system maintenance [8,14-18].

In addition, a study evaluating the main factors of failure in implementing HISs in several African hospitals showed that the main difficulties encountered were not technical [8]. They were mainly human, cultural, social, administrative, and environmental, such as skepticism, resistance to change, insufficient ICT skills, poor organization (operating in silos), unrealistic implementation deadlines, insufficient technical support, and lack of support for users after implementation [8,17].

On the other hand, factors strongly associated with success were adequate allocation of resources (infrastructure, human, and financial), good communication, good planning and project management, ability to reorganize the institution to adapt to new processes, local buy-in from stakeholders, implementation based on a holistic approach (integration of all services), highlighting successes to motivate users and reduce mistrust, adequate technical support, perceived usefulness and user satisfaction [6,16,19,20]. To increase the chances of success, in addition to technical aspects (software and equipment) and human resources, good planning, stakeholder commitment, and in-depth change management were required [7]. It is also essential to prioritize design: “enhancing consultations during the intervention design, better consideration of implementation challenges during the design and better recognition of relations between different influences” [21].

By taking into account lessons learned from previous CIS or HIS implementations, including those conducted in resource-limited settings and in addition to technical aspects, we emphasized the management, organizational, leadership, and change management aspects during the project’s implementation. This concerned the representation of all entities (steering committee, project team, and user group), phase-based project management, marketing of the project, management of communication with all stakeholders, drafting and implementation of a procedure manual, use of local champions, change management activities (user awareness–raising activities and allocation of dedicated human resources for responsive users’ support), and finally, the training of users. We also sent a letter and had a meeting with the director of the YCH about the project and raised awareness among health care professionals, managers, and decision makers about the importance of an HIS or CIS in a hospital through scientific conferences and the general public media, etc.

Despite all these approaches, and as shown by our results, beyond the project management, technical, and financial aspects, the other main problems of implementing health or hospital information systems in sub-Saharan Africa lies in digital health leadership, governance, and change management. This digital health leadership, governance, and change management should prioritize data as a tool for improving productivity and managing health care institutions, promote a data culture among health care professionals to support the change of mindset, and the acquisition of information management skills (collection, processing, discussion, and use).

However, the use of good practice in managing a digital health project for setting up a new system by incorporating a good grasp of the project management, technical, organizational, leadership and cultural aspects does not always guarantee its success [18]. In countries with a highly centralized political system such as ours, it is very often necessary for such projects to have a high-level strategic and political anchor (for example, at the level of the ministries responsible for public health or telecommunications or even higher authorities such as the prime minister or the president of the republic). The government has a critical role in developing a vision and creating the foundation upon which innovation activities will be developed [22]. In Mali, for example, the success of the implementation of the HIS at the CHU Mère-Enfant Le Luxembourg was partially based on the support received by the country's first lady. Furthermore, a decision by the minister of health in 2011, with the advent of health insurance in Mali, enabled this tool to be deployed in 72 hospitals nationwide. In the case of our project, at the time of its implementation, Cameroon did not yet have a vision or strategy for digital health. The strategy was adopted in January 2020 and contains 7 strategic objectives covering governance and leadership, legal and regulatory framework, human resources, funding and investment, services and applications, ICT infrastructure, and standards and interoperability [23]. Even if it is still in its embryonic stages, we note that since the adoption of this strategy and following the outbreak of the COVID-19 pandemic, the State of Cameroon is gradually pushing for the adoption and integration of digital health tools

into our health system. An operational plan linked to this strategic plan is currently being designed and validated.

Acknowledgments

This project was implemented with the agreement of both partners Yaoundé Central Hospital and Geneva University Hospitals.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Checklist of iCHECK-DH guidelines. iCHECK-DH: Guidelines and Checklist for the Reporting on Digital Health Implementations.
[PDF File (Adobe PDF File), 146 KB - medinform_v11i1e48256_app1.pdf]

References

1. Everybody's business -- strengthening health systems to improve health outcomes. World Health Organization. 2007. URL: https://www.who.int/publications/i/item/everybody-s-business----strengthening-health-systems-to-improve-health-outcomes [accessed 2023-10-02]
2. Ngwakongnwi E, Atanga MBS, Quan H. Challenges to implementing a national health information system in Cameroon: perspectives of stakeholders. J Public Health Afr 2014 Feb 04;5(1):322 [FREE Full text] [doi: 10.4081/jphia.2014.322] [Medline: 28299116]
3. Classification of digital health interventions v1.0. World Health Organization. URL: https://apps.who.int/iris/bitstream/handle/10665/260480/WHO-RHR-18.06-eng.pdf [accessed 2023-02-15]
4. Marutha NS, Ngoepe M. The role of medical records in the provision of public healthcare services in the Limpopo province of South Africa. S Afr J Inf Manag 2017 Sep 27;19(1). [doi: 10.4102/sajim.v19i1.873]
5. Amarasingham R, Plantinga L, Diener-West M, Gaskin DJ, Powe NR. Clinical information technologies and inpatient outcomes: a multiple hospital study. Arch Intern Med 2009 Jan 26;169(2):108-114. [doi: 10.1001/archinternmed.2008.520] [Medline: 19171805]
6. Bagayoko C, Dufour J, Chaacho S, Bouhaddou O, Fieschi M. Open source challenges for hospital information system (HIS) in developing countries: a pilot project in Mali. BMC Med Inform Decis Mak 2010 Apr 16;10(1):22 [FREE Full text] [doi: 10.1186/1472-6947-10-22] [Medline: 20398366]
7. Madore A, Rosenberg J, Muyindike WR, Bangsberg DR, Bwana MB, Martin JN, et al. Implementation of electronic medical records requires more than new software: Lessons on integrating and managing health technologies from Mbarara, Uganda. Healthc (Amst) 2015 Dec;3(4):264-269. [doi: 10.1016/j.hjdsi.2015.08.006] [Medline: 26699355]
8. Verbeke F, Karara G, Nyssen M. Human factors predicting failure and success in hospital information system implementations in sub-Saharan Africa. Stud Health Technol Inform 2015;216:482-486. [Medline: 26262097]
9. Perrin Franck C, Babington-Ashaye A, Dietrich D, Bediang G, Veltsos P, Gupta PP, et al. iCHECK-DH: Guidelines and Checklist for the Reporting on Digital Health Implementations. J Med Internet Res 2023 May 10;25:e46694 [FREE Full text] [doi: 10.2196/46694] [Medline: 37163336]
10. MEDIBOARD. URL: https://mediboard.org/ [accessed 2023-09-01]
11. Method overview. HERMES Online. URL: https://www.hermes.admin.ch/en/project-management/understanding/overview-hermes/method-overview.html [accessed 2022-09-01]
12. Bagayoko CO. Mise en place d'un Système d'Information Hospitalier en Afrique Francophone : Cinz@n, étude et validation du modèle au Mali. Université de la Méditerranée. 2010. URL: https://www.theses.fr/2010AIX20680.pdf [accessed 2023-10-02]
13. Wendt T, Häber A, Brigl B, Winter A. Modeling hospital information systems (Part 2): using the 3LGM2 tool for modeling patient record management. Methods Inf Med 2018 Feb 05;43(03):256-267. [doi: 10.1055/s-0038-1633866]
14. Muinga N, Magare S, Monda J, English M, Fraser H, Powell J, et al. Digital health systems in Kenyan public hospitals: a mixed-methods survey. BMC Med Inform Decis Mak 2020 Jan 06;20(1):2 [FREE Full text] [doi: 10.1186/s12911-019-1005-7] [Medline: 31906932]
15. Mihalas GI. Analysis of barriers in implementation of health information systems - EFMI conference introductory address. Stud Health Technol Inform 2008;134:21-26. [Medline: 18376030]
16. Hanmer LA, Isaacs S, Roode JD. A conceptual model of computerised hospital information system (CHIS) use in South Africa. Stud Health Technol Inform 2007;129(Pt 1):63-67. [Medline: 17911679]
17. Muinga N, Magare S, Monda J, Kamau O, Houston S, Fraser H, et al. Implementing an open source electronic health record system in Kenyan health care facilities: case study. JMIR Med Inform 2018 Apr 18;6(2):e22 [FREE Full text] [doi: 10.2196/medinform.8403] [Medline: 29669709]

18. Kpobi L, Swartz L, Ofori-Atta AL. Challenges in the use of the mental health information system in a resource-limited setting: lessons from Ghana. BMC Health Serv Res 2018 Feb 08;18(1):98 [FREE Full text] [doi: 10.1186/s12913-018-2887-2] [Medline: 29422047]
19. Bisrat A, Minda D, Assamnew B, Abebe B, Abegaz T. Implementation challenges and perception of care providers on Electronic Medical Records at St. Paul's and Ayder Hospitals, Ethiopia. BMC Med Inform Decis Mak 2021 Nov 02;21(1):306 [FREE Full text] [doi: 10.1186/s12911-021-01670-z] [Medline: 34727948]
20. Gannon H, Chimhuya S, Chimhini G, Neal SR, Shaw LP, Crehan C, et al. Electronic application to improve management of infections in low-income neonatal units: pilot implementation of the NeoTree beta app in a public sector hospital in Zimbabwe. BMJ Open Qual 2021 Jan 20;10(1):e001043 [FREE Full text] [doi: 10.1136/bmjoq-2020-001043] [Medline: 33472853]
21. Ahuja S, Mirzoev T, Lund C, Ofori-Atta A, Skeen S, Kufuor A. Key influences in the design and implementation of mental health information systems in Ghana and South Africa. Glob Ment Health (Camb) 2016 Apr 08;3:e11 [FREE Full text] [doi: 10.1017/gmh.2016.3] [Medline: 28596880]
22. Desveaux L, Soobiah C, Bhatia RS, Shaw J. Identifying and overcoming policy-level barriers to the implementation of digital health innovation: qualitative study. J Med Internet Res 2019 Dec 20;21(12):e14994 [FREE Full text] [doi: 10.2196/14994] [Medline: 31859679]
23. Plan Stratégique National de Santé Numérique 2020–2024. Ministère de la Santé Publique du Cameroun. 2019. URL: https://tinyurl.com/nye6wbfj [accessed 2023-07-13]

Abbreviations

CCAM: Classification Commune des Actes Médicaux
CIS: clinical information system
HIS: hospital information system
HL7: Health Level 7
HPRIM: harmoniser et promouvoir l'informatique médicale
HUG: Geneva University Hospitals
***ICD-10*:** *10th revision of the International Classification of Diseases*
iCHECK-DH: Guidelines and Checklist for the Reporting on Digital Health Implementations
ICT: information and communication technology
YCH: Yaoundé Central Hospital

Edited by C Lovis, C Perrin; submitted 17.04.23; peer-reviewed by A Koumamba, A Babington-Ashaye; comments to author 23.06.23; revised version received 25.07.23; accepted 26.08.23; published 18.10.23.

Please cite as:
Bediang G
Implementing Clinical Information Systems in Sub-Saharan Africa: Report and Lessons Learned From the MatLook Project in Cameroon
JMIR Med Inform 2023;11:e48256
URL: https://medinform.jmir.org/2023/1/e48256
doi:10.2196/48256
PMID:37851502

Implementation Report

Clinical Decision Support to Reduce Opioid Prescriptions for Dental Extractions using SMART on FHIR: Implementation Report

D Brad Rindal[1*], DDS; Dhavan Prasad Pasumarthi[1*], BTech, CSIT; Vijayakumar Thirumalai[1*], BTech; Anjali R Truitt[2*], MPH, PhD; Stephen E Asche[1], MA; Donald C Worley[1], DDS; Sheryl M Kane[1], BS; Jan Gryczynski[3], PhD; Shannon G Mitchell[3], PhD

[1]HealthPartners Institute, Minneapolis, MN, United States

[2]Memorial Hermann Health System, Houston, TX, United States

[3]Friends Research Institute, Baltimore, MD, United States

[*]these authors contributed equally

Corresponding Author:
D Brad Rindal, DDS
HealthPartners Institute
8170 33rd Ave S
Minneapolis, MN, 55425
United States
Phone: 1 952 967 5026
Email: donald.b.rindal@healthpartners.com

Abstract

Background: Clinical decision support (CDS) has the potential to improve clinical decision-making consistent with evidence-based care. CDS can be designed to save health care providers time and help them provide safe and personalized analgesic prescribing.

Objective: The aim of this report is to describe the development of a CDS system designed to provide dentists with personalized pain management recommendations to reduce opioid prescribing following extractions. The use of CDS is also examined.

Methods: This study was conducted in HealthPartners, which uses an electronic health record (EHR) system that integrates both medical and dental information upon which the CDS application was developed based on SMART (Substitutable Medical Applications and Reusable Technologies) on FHIR (Fast Healthcare Interoperability Resources). The various tools used to bring relevant medical conditions, medications, patient history, and other relevant data into the CDS interface are described. The CDS application runs a drug interaction algorithm developed by our organization and provides patient-specific recommendations. The CDS included access to the state Prescription Monitoring Program database.

Implementation (Results): The pain management CDS was implemented as part of a study examining opioid prescribing among patients undergoing dental extraction procedures from February 17, 2020, to May 14, 2021. Provider-level use of CDS at extraction encounters ranged from 0% to 87.4% with 12.1% of providers opening the CDS for no encounters, 39.4% opening the CDS for 1%-20% of encounters, 36.4% opening it for 21%-50% of encounters, and 12.1% opening it for 51%-87% of encounters.

Conclusions: The pain management CDS is an EHR-embedded, provider-facing tool to help dentists make personalized pain management recommendations following dental extractions. The SMART on FHIR–based pain management CDS adapted well to the point-of-care dental setting and led to the design of a scalable CDS tool that is EHR vendor agnostic.

Trial Registration: ClinicalTrials.gov NCT03584789; https://clinicaltrials.gov/study/NCT03584789

(JMIR Med Inform 2023;11:e45636) doi:10.2196/45636

KEYWORDS

clinical decision support systems; dentistry; analgesics; electronic health records; EHR; algorithm; design; implementation; decision support; development; dentists; pain management; patient care; application; tool; Fast Healthcare Interoperability Resources; FHIR; Substitutable Medical Applications and Reusable Technologies; SMART

Introduction

The United States has experienced an epidemic of opioid overdose deaths, with deaths associated with prescription pain relievers of particular concern [1]. Inappropriate prescribing of opioids, heroin use, and the increase in the use of illicitly manufactured fentanyl and its analogues have driven this unprecedented opioid epidemic. Opioid analgesics are among the most frequently prescribed drugs by dentists [2]. An estimated 5 million people undergo third-molar extractions in the United States each year [3]. Evidence shows that exposure to opioid analgesic prescriptions following dental extractions and other procedures is widespread in the United States [4]. Combining a nonsteroidal anti-inflammatory drug with acetaminophen provides a viable and evidence-based pain management alternative to prescription opioids when better pain control is needed [5].

Electronic health records (EHRs) may contain much of the relevant medical history information needed to make appropriate decisions without navigation of multiple screens required to locate desired information. Unfortunately, dental EHR systems are generally not part of the medical EHRs; therefore, dentists rely on the patients to complete a medical history questionnaire. Currently, the information exchange between dentistry and medicine is hampered by a lack of data standards and interoperability between medical and dental EHR systems [6].

Clinical decision support (CDS) has the potential to improve clinical decision-making consistent with evidence-based care [7-9]. CDS can be designed to save providers time and help them provide safe, personalized analgesic prescribing by bringing together relevant medical conditions, current medications, a prior history of substance use, and additional prescribing information from the state prescription drug monitoring program. The problem with integrating siloed yet important health information [6] and the proposed solution of using CDS to integrate this information into one interface serves as the premise for this research study [10].

The objective of this study was to test the efficacy of two interventions (CDS with and without patient education), compared to the treatment-as-usual approach to decrease opioid prescribing for dental extractions. This manuscript adheres to the iCHECK-DH (Guidelines and Checklist for the Reporting on Digital Health Implementations) [11]. The National Institute of Dental and Craniofacial Research of the National Institutes of Health (NIH) funded this project using a cooperative agreement where they provided oversight, coordination, and facilitation. This paper describes the development of a CDS system using HL7 (Health Level 7) FHIR (Fast Healthcare Interoperability Resources) and the SMART (Substitutable Medical Applications and Reusable Technologies) on FHIR framework.

Methods

Development of the CDS

The project overview is presented in Multimedia Appendix 1.

The CDS is built on secure, scalable, and EHR-agnostic core design principles using health care systems' critical IT infrastructure. Point-of-care, real-time patient medical information extraction is key to the CDS system. This allows for accurate personalized recommendations and integration with the EHR in a manner that fits into the clinical workflow without deviating from the standard-of-care process, facilitating improved CDS use with low or no burden on the health care provider. The CDS system used the industry-standard EHR interoperability method, HL7 FHIR, for data formats and the application programming interface (API) for data exchanges and health care IT. The EHR-agnostic third-party application integration method, the "SMART on FHIR" framework, provides secure authentication and authorization management to the application through a valid existing EHR session. The SMART on FHIR framework built on OAuth 2.0 [12] authorization framework provides secure token and code exchanges to establish access to the FHIR resource server for patient medical record data extraction and presents clinical recommendations in a dental provider interface.

The DIODE ("De-Implementing Opioid Use and Implementing Optimal Pain Management Following Dental Extractions" study) CDS tool follows the EHR data governance model to protect the data privacy and data security of patients. The backend database system for the DIODE CDS tool follows the highest industry standards and enterprise access control policies, limiting data access to only authorized IT personnel who are designated for maintaining the systems. The SMART on FHIR–based DIODE CDS application design approach follows EHR access control policies and privacy settings, allowing only authorized individuals with valid HER-authenticated user sessions to access patient records.

Translating Local Data Into FHIR-Compliant Data

Data access is limited to specific FHIR resources (Table 1) through the OAuth 2.0 authorization scope, and API access expires in a short time for improved data privacy and security. The pain management CDS is only launched from a valid EHR session on specific patient records, based on context provided by the dentist in a secure manner, enabling the pain management CDS as a single sign-on and saving time for the provider.

Since different versions of HL7 FHIR specification have been developed and implemented by EHR vendors, the pain management CDS selected the HL7 FHIR DSTU2 version, which is supported by multiple leading EHR systems.

Table 1. FHIR (Fast Healthcare Interoperability Resources) resource and use.

Source	FHIR resource	Pain management CDS[a] use
1	AllergyIntolerance.Search (DSTU2)	Relevant allergies.
2	Condition.Search (DSTU2)	Relevant medical conditions from the active problem list and diagnosis.
3	MedicationStatement.Search (DSTU2)	Current medications with potential interactions with pain medications commonly prescribed by dentists.
4	Observation.Search (DSTU2)	Nonprescription substances used by the patient.
5	Patient.Read (DSTU2)	Relevant patient demographic information.

[a]CDS: clinical decision support.

Pain Medication Recommendations (Algorithm Development)

Pertinent patient medical information is extracted from the EHR as FHIR resources. The pain management CDS uses an algorithm developed by our team to identify patient-specific contraindications for pain medications, including opioid medications, nonsteroidal anti-inflammatory drugs, and acetaminophen. It also identifies patient allergies and conditions potentially impacted by these analgesics.

RxNorm API

RxNav is a service of the National Library of Medicine, which is part of the NIH, an agency of the US Department of Health and Human Services [13]. RxNav is a browser for several drug information sources, including RxNorm, RxTerms, and Medication Reference Terminology (MED-RT). We relied on the RxNorm code system to translate the system-specific drug or medication names into standardized names. These codes are further used to get ingredient-level information using the RxNav API web service [14]. The ingredient-level information is used in a custom-defined algorithm to find the drug-level interactions.

Medication ingredient-level information is incorporated into the RxNorm web service API to convert brand name or generic name drugs from the EHR to normalized RxNorm code details. The medication ingredient-level information is used in an in-house custom-defined algorithm to identify drug-level interactions. The pain management CDS application uses these drug-level interactions to display patient-level personalized pain management recommendations.

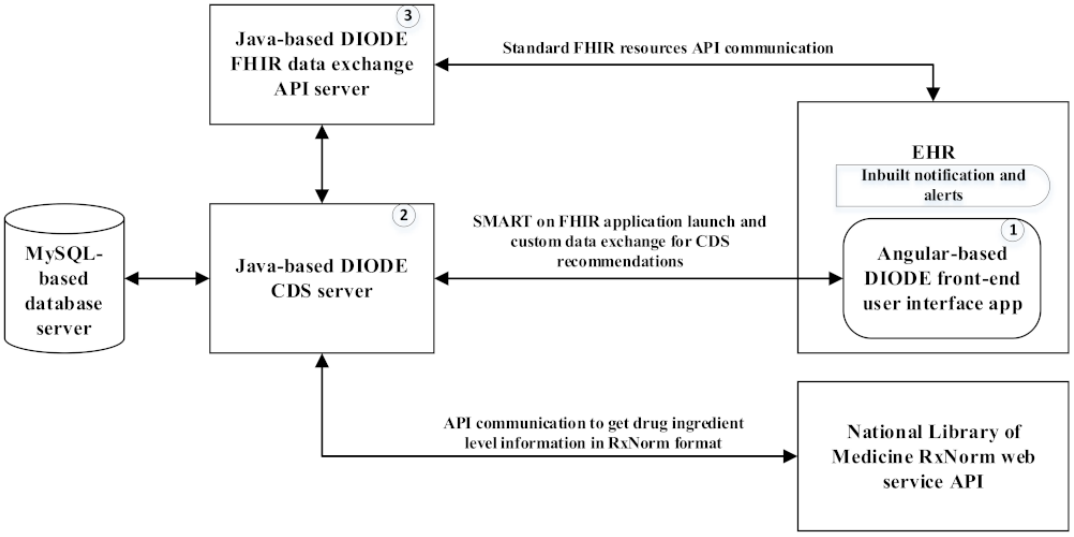
EHR Integration

The CDS is designed as a software component-based, reusable subsystem to perform various functions and data flow support. Angular was used as the framework for the user interface, which was embedded into the EHR for integrated workflow support. The Java-based DIODE CDS server is the heart of the system, which performs algorithm processing, data storage, and retrieval from the database server and provides personalized clinical recommendations. The Java-based DIODE FHIR data exchange server facilitates FHIR resource exchanges with the EHR FHIR resource server and provides collected data to the DIODE CDS server. Figure 1 explains the application architecture.

The Angular-based DIODE front-end application was developed for presenting CDS personalized recommendations to an intuitive user interface with subsections for (1) drug interactions, (2) condition considerations, as well as (3) allergies or intolerances and relevant action items. The front-end application is launched from the EHR and communicates with the DIODE CDS server to get the required data to generate the user interface and education materials.

The CDS server is the core of the pain management CDS, which connects to other software components for data extraction and processing, including the management of OAuth 2.0 communication with the EHR authorization server. The CDS server connects to the backend database server for storing patient data extracted from the EHR, which includes personalized recommendation information generated by CDS algorithms. It also retrieves data from the database server to generate the user interface. The FHIR exchange API server communicates with the EHR FHIR resource server to extract patient medical record information in JavaScript Object Notation standard text-based format. The HER's inbuilt Notification and Alert functionality highlights the CDS prompt in the record when an appropriate clinical condition is met, such as the addition of a tooth extraction procedure to the patient's chart. By default, the dentist has access to the pain management CDS from the patient header section, and only eligible patient records show the highlighted link to the CDS. A MySQL database server is used to store the CDS data, which includes extracted patient clinical data from the EHR, algorithm mapping data, and generated personalized CDS recommendations.

Figure 1. The pain management clinical decision support system (CDS) application architecture. API: application programming; CDS: clinical decision support; DIODE: De-Implementing Opioid Use and Implementing Optimal Pain Management Following Dental Extractions; FHIR: Fast Healthcare Interoperability Resources interface; SMART: Substitutable Medical Applications and Reusable Technologies.



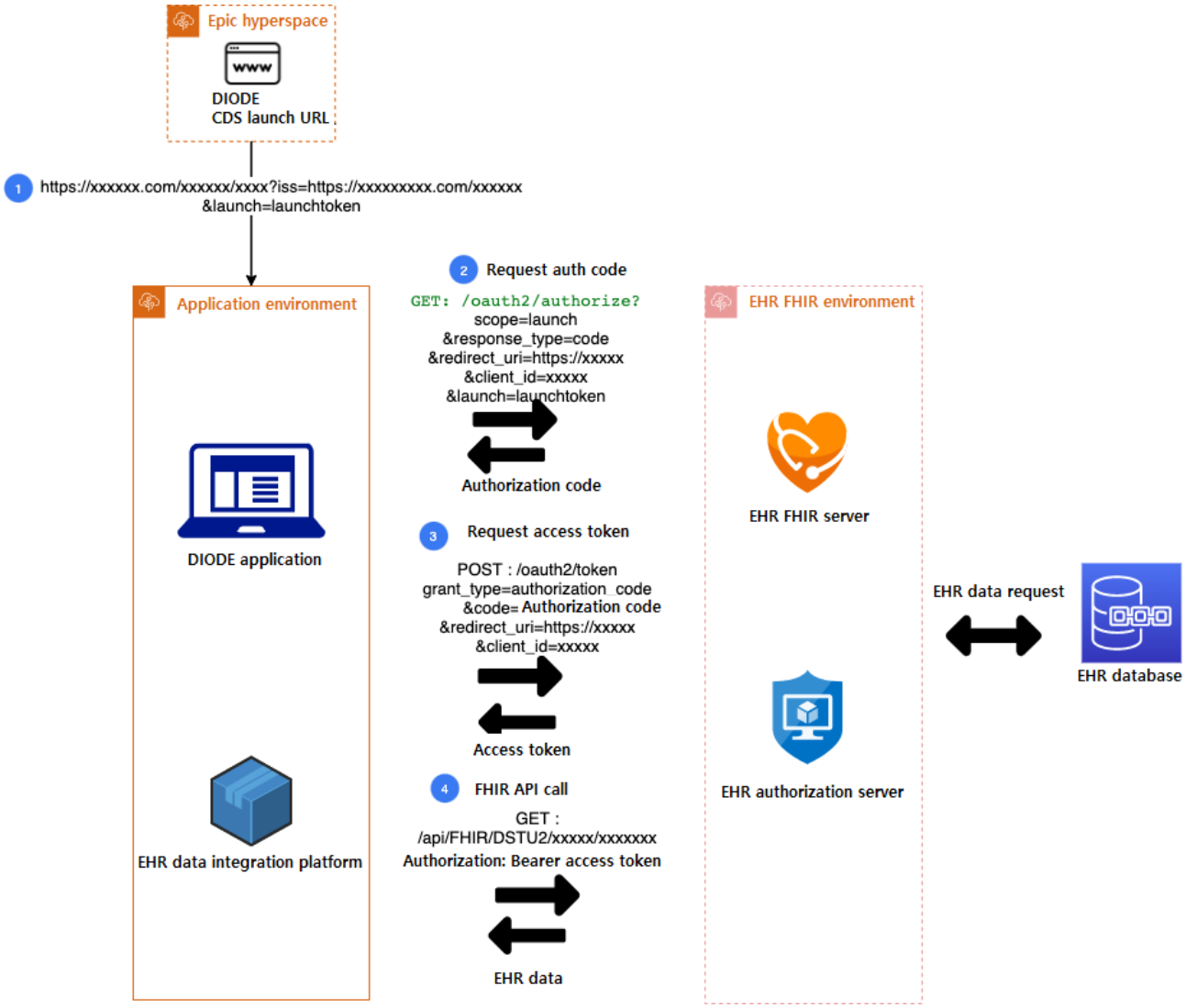
Access to the Minnesota Prescription Monitoring Program

The CDS provides a hyperlink to the Minnesota Prescription Monitoring Program for providers to view controlled substance prescriptions from the state-level prescription monitoring database.

Launching the CDS Application

Since the CDS is embedded in the EHR, the application can be launched, when needed, using a navigation link integrated within the EHR. In the initial launch, the application receives a launch token, which provides a valid EHR user session context and a URL; the URL helps to identify EHR FHIR server information, including EHR authorization server and resource server URL end points. The application receives the authorization code after a successful request along with a launch token. The authorization code prompts an access token, which is used subsequently in multiple FHIR resource API call requests to get the required patient information for CDS use. In the EHR, a header alert is highlighted when patient eligibility criteria are met. The CDS can be launched with 1 click, simplifying navigation to the CDS. Additionally, an HER-specific navigational link was created to access the pain management CDS from any patient chart context. Figure 2 [12] explains the data extraction flow.

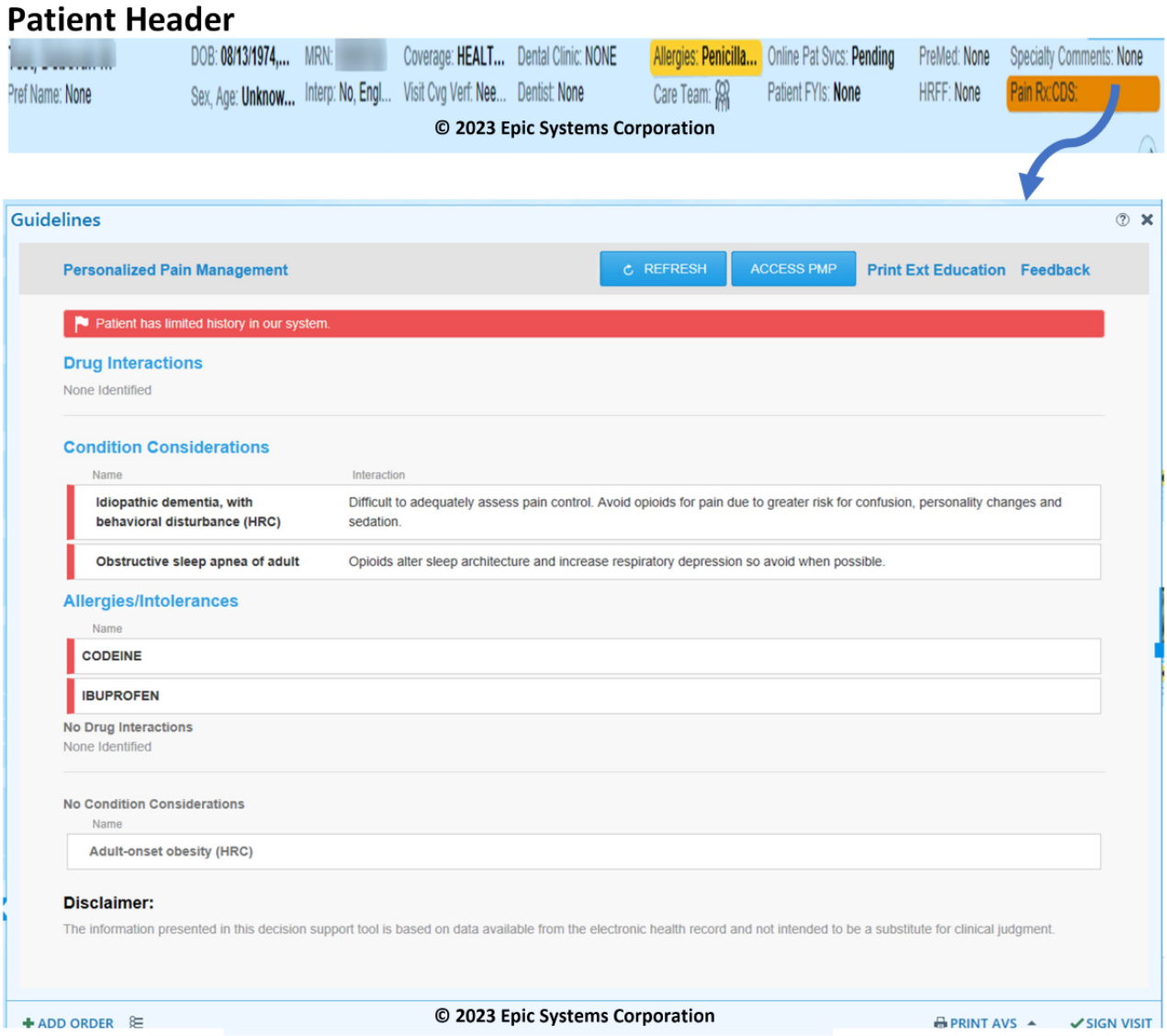
Figure 2. The pain management—OAuth2.0 [12] and the HL7 (Health Level 7) FHIR (Fast Healthcare Interoperability Resources) resource data extraction dataflow diagram. CDS: clinical decision support. DIODE: De-Implementing Opioid Use and Implementing Optimal Pain Management Following Dental Extractions; EHR: electronic health record; FHIR: Fast Healthcare Interoperability Resources.



Pain Management CDS Application Interface

The CDS interface provides a simple summary (Figure 3) of the relevant information that should be considered in deciding about the most appropriate pain management strategy, including the most appropriate analgesic to prescribe or recommend for an individual patient. Relevant information was highlighted in red.

Figure 3. Clinical decision support (CDS) system screen example.



Addressing Missing HL7 FHIR Resources

The CDS extracts data from the EHR using the HL7 FHIR DSTU2 standard data model, which was lacking a few data items at the time of the CDS design. A few custom data elements were added, such as (1) Patient Pregnancy Status, (2) Patient Breastfeeding Status, and (3) Substance Use Status. These data were extracted using custom functionality provided by the EHR system (Epic Extensions) and sent as part of the SMART on FHIR access token response, which is an extension of standard OAuth 2.0 token response after successful authorization validation.

Addressing Other Technical Requirements

AngularJS is a JavaScript-based open-source front-end web framework for developing single-page applications. It is maintained mainly by Google and a community of individuals and corporations. It aims to simplify both the development and the testing of such applications by providing a framework for client-side model–view–controller and model–view–view–model architectures, along with components commonly used in web applications and progressive web applications. AngularJS implements the model–view–controller pattern to separate presentation, data, and logic components [15]. Using dependency injection, Angular brings traditionally server-side services, such as view-dependent controllers, to client-side web applications. Consequently, much of the burden on the server can be reduced. AngularJS brings value when dealing with non-Epic systems.

Statistical Analysis of CDS Use

The CDS was highlighted to alert the dental provider when an extraction was planned but did not open automatically. Therefore, this design offered an opportunity to measure when a clinician opened the CDS, which was an important outcome of CDS use.

Dental provider attributes and the frequency of CDS use are described with counts, percentages, and means (SDs). Differences in CDS use by provider attributes, including provider sex, provider age (<40 vs >40 years), and number of extraction encounters (<100 vs >100) are tested in a generalized linear mixed model (with a logit link and binomial error distribution) containing fixed effects for the provider attributes

and a random intercept for the provider to accommodate the clustering of patient encounters within providers. Model-derived percentages and *P* values are presented.

Implementation (Results)

Among 20 clinics with providers assigned to the intervention arms, 95.0% (19/20) of clinics had at least 1 extraction encounter with the CDS opened. Among 3 oral surgeons assigned to the active intervention arms, 2 opened the CDS for 0.5% (10/1874) of tooth extraction encounters. Among 30 dentists assigned to the active intervention arms, 27 opened the CDS for 24.3% (501/2059) of tooth extraction encounters. Provider-level use of the CDS at extraction encounters ranged from 0% to 87.4%, with 12.1% of providers never opening the CDS (0% of encounters), 39.4% opening the CDS for 1%-20% of encounters, 36.4% opening it for 21%-50% of encounters, and 12.1% opening it for 51%-87% of encounters.

Among 2059 encounters linked to 30 dentists in the active intervention arms, the CDS was opened at similar levels by dentists aged 40 years and younger compared to dentists older than 40 years (27.0% vs 22.5%; *P*=.62). Male and female dentists opened the CDS at similar levels (24.0% vs 26.0%; *P*=.83). Dental providers with 100 or more extraction encounters opened the CDS at similar levels as those with fewer than 100 extraction encounters (26.1% vs 21.5%; *P*=.62).

Among 1061 extraction encounters in the two intervention arms in which an opioid was prescribed, the CDS was opened for 5.4% (57/1061) of encounters. Among 2872 extraction encounters in the two intervention arms in which an opioid was not prescribed, the CDS was opened for 15.8% (454/2872) of encounters.

Ethical Considerations

The study was approved by the HealthPartners Institutional Review Board (#A17-013). The project was part of a broader quality improvement initiative approved by the HealthPartners Dental Group and did not alter the standard of care for dental extractions; approval by the Dental Group for this minimal risk study was acceptable as an alternative to written informed consent documentation for dentists. Patients on the research exclusion list were not included in the study.

Discussion

Principal Findings

This study used a programming interface and integration platform that combines with existing EHRs, patient portals, personal health records, and data warehouses. The 3 key aspects of SMART on FHIR are as follows: (1) a data access layer based on FHIR, combined with a set of constraining profiles that lock down optionality and align vocabularies with Meaningful Use requirements; (2) a security layer that provides narrowly scoped authorization to specific portions of a patient's record via OAuth 2.0; and (3) a single-sign-on layer using OpenID Connect, which can either integrate with an existing EHR or patient portal session, conveying the current patient, encounter, and other host environment details, or launch independently, such as on a mobile phone or device [12,16,17].

The HL7 FHIR standards are constantly improved, and new data items are added to FHIR resources data model with each new version [18]. Since the custom data extraction varies for each EHR vendor, some data items from the patient's medical record may not be part of the standard FHIR resources. The custom data elements are sent as part of Access Token Response after successful Authorization providing streamlined data flow without comprising IT security.

The CDS application link is built into the patient header section in the EHR as a clickable link, which can be launched anytime as a SMART on FHIR application by a dentist. The clickable link is highlighted in a yellow background color to get the provider's attention when an eligible criterion (ie, a tooth extraction procedure) is added to the patient's chart. This approach is different from the HL7 CDS Hooks approach, which sends the information to the CDS server system on a specific event in the patient's chart, such as chart open or order entry, and provides a response to show actionable CDS card information.

Overall, results showed that health care providers' use of the pain management CDS was low and consistent with CDS use in other studies [19]. Postintervention interviews with providers indicated that the app worked exactly as developed and that those who used it regularly found the synthesized health information to be very beneficial, informative, and time saving. Providers who did not use it more than once or twice identified several reasons for not using it consistently. These reasons included the following: forgetting about the CDS, which is plausible, as it was not triggered unless an extraction was treatment planned; some dental providers not modifying their workflow to include checking the CDS; and some finding the visual representation (highlighting in the EHR) to be subtle and easily overlooked in a busy EHR dashboard. Early attempts at opening the CDS that resulted in either problems with functionality or providers determining that the type of information they expected to see was not included also negatively impacted its continued use.

CDS strategies to deimplement opioid prescribing for dental extractions did not lead to reduced opioid prescribing compared to standard practice in the main trial. We found that opioid prescribing declined significantly over time in all conditions. A comprehensive description and results of this clinical trial have been published [20]. HealthPartners [21] was already paying close attention to opioids and undertaking several actions to reduce opioid prescribing by providers while the study was being implemented. In the context of a downward trend in opioid prescribing, dental providers identified several factors that led to reduced reliance on opioids, including governmental and health system opioid prescribing policy changes and the COVID-19 pandemic [22].

We trained providers to open the CDS for all extraction procedures so they could receive a summary of relevant information about potential drug interactions and medical conditions relevant to analgesic prescribing. In qualitative interviews conducted at the end of the study, dentists often described the CDS as something they perceived as useful only if they were considering prescribing an opioid. If another

analgesic option was planned or the patient did not have any medical conditions or few current medications, the providers did not see a reason to open the CDS [22]. These findings suggest that the appropriate use of CDS needs to be tailored to more complex or high-risk patient groups. In addition, the training information was sent in an email communication that included training slides and continuing education credit for viewing the slides. We provided contact information for a study team member if there were any questions. CDS was likely negatively impacted by a passive approach to email communication. A more robust in-person or web-based training session for the providers to attend could improve CDS use.

CDS application using the SMART on FHIR framework improves health care interoperability similar to standards applied to other industries [23]. Most EHR databases use a proprietary API. Without tools like SMART on FHIR, it becomes necessary to build a custom connection to each database to access medical data. This is costly, hindering the ability of health care providers and patients to access their data with their preferred technology [24]. SMART on FHIR provides a standard, universal API for accessing the EHR database they use. We estimate that 20% of the budget was used to build and integrate various data elements into the CDS tool. This research project was budgeted at US $500,000 a year and funded for 5 years without an option for additional funding.

Conclusions

Development of the provider-facing, point-of-care CDS using SMART on FHIR supported the study goals of providing dental practitioners with context-dependent pain management alternatives and enabling patient-centered shared decision-making for pain management. However, CDS use was limited. Strategies to improve its use need to be considered for CDS tools to realize their potential. Future opportunities for CDS use and tools such as SMART on FHIR include clinical topics where provider behavior needs to align with current evidence and would benefit from integrating medical data. Table 2 summarizes key insights and recommendations. This report follows the Implementation Reporting Guidelines.

Table 2. Insights and recommendations.

Insight	Recommendation
Well-designed CDS[a] does not insure utilization.	Engage providers in the design and provide robust training.
Linking data from an EHR[b] into a CDS does not require the use of tools such as SMART[c] on FHIR[d].	SMART on FHIR tools allow for interoperability.
Not all well-intended provider-focused CDS use results in improved care.	Test the CDS to determine if it achieves the desired change.

[a]CDS: clinical decision support.

[b]EHR: electronic health record.

[c]SMART: Substitutable Medical Applications and Reusable Technologies.

[d]FHIR: Fast Healthcare Interoperability Resources.

Acknowledgments

Research reported in this publication was supported by the National Institute of Dental and Craniofacial Research of the National Institutes of Health (NIH) under (Award U01DE027441). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Authors' Contributions

All authors were involved in conducting the study. DPP and VT were responsible for building the SMART on FHIR application. DPP, VT, ART, SEA, and DBR were involved in drafting the manuscript. SMK had full access to all the data in the study. SEA worked with SMK to ensure data accuracy and was responsible for data analysis. All authors critically reviewed and approved the manuscript. SGM and DBR were the study’s principal investigators.

Conflicts of Interest

Unrelated to this study, JG is an investigator on an NIH-funded study that has received free medication from Indivior and Alkermes, has received research funding from Indivior (paid to his institution and including project-related salary support) and has part of the ownership of COG Analytics. The other authors have no potential conflicts of interest to report.

Multimedia Appendix 1
Clinical decision support system (CDS) development overview.
[PPTX File , 37 KB - medinform_v11i1e45636_app1.pptx]

References

1. Report to congressional requesters. Prescription pain reliever abuse. United States Government Accountability Office. 2011 Dec. URL: https://www.gao.gov/assets/gao-12-115.pdf [accessed 2020-04-26]
2. Moore PA, Nahouraii HS, Zovko JG, Wisniewski SR. Dental therapeutic practice patterns in the U.S. II. Analgesics, corticosteroids, and antibiotics. Gen Dent 2006;54(3):201-7; quiz 208, 221. [Medline: 16776415]
3. Friedman JW. The prophylactic extraction of third molars: a public health hazard. Am J Public Health 2007 Sep;97(9):1554-1559. [doi: 10.2105/AJPH.2006.100271] [Medline: 17666691]
4. Steinmetz CN, Zheng C, Okunseri E, Szabo A, Okunseri C. Opioid analgesic prescribing practices of dental professionals in the United States. JDR Clin Trans Res 2017 Jul;2(3):241-248 [FREE Full text] [doi: 10.1177/2380084417693826] [Medline: 28879246]
5. Moore PA, Hersh EV. Combining ibuprofen and acetaminophen for acute pain management after third-molar extractions: translating clinical research to dental practice. J Am Dent Assoc 2013 Aug;144(8):898-908. [doi: 10.14219/jada.archive.2013.0207] [Medline: 23904576]
6. Rajkumar N, Muzoora M, Thun S. Dentistry and interoperability. J Dent Res 2022 Oct 10;101(11):1258-1262 [FREE Full text] [doi: 10.1177/00220345221100175] [Medline: 35689387]
7. O'Connor PJ, Sperl-Hillen JM. Outpatient clinical decision systems that work: lessons learned from research and experience. 2020 Presented at: Duke University Grand Rounds; Oct 23; Virtual event.
8. Bates DW, Kuperman GJ, Wang S, Gandhi T, Kittler A, Volk L, et al. Ten commandments for effective clinical decision support: making the practice of evidence-based medicine a reality. J Am Med Inform Assoc 2003;10(6):523-530 [FREE Full text] [doi: 10.1197/jamia.M1370] [Medline: 12925543]
9. Sperl-Hillen JM, Rossom RC, Kharbanda EO, Gold R, Geissal ED, Elliott TE, et al. Priorities wizard: multisite web-based primary care clinical decision support improved chronic care outcomes with high use rates and high clinician satisfaction rates. EGEMS (Wash DC) 2019 Apr 03;7(1):9. [doi: 10.5334/egems.284] [Medline: 30972358]
10. Jung S, Bae S, Seong D, Oh OH, Kim Y, Yi B. Shared interoperable clinical decision support service for drug-allergy interaction checks: implementation study. JMIR Med Inform 2022 Nov 10;10(11):e40338 [FREE Full text] [doi: 10.2196/40338] [Medline: 36355401]
11. Perrin Franck C, Babington-Ashaye A, Dietrich D, Bediang G, Veltsos P, Gupta PP, et al. iCHECK-DH: Guidelines and Checklist for the Reporting on Digital Health Implementations. J Med Internet Res 2023 May 10;25(4):e46694 [FREE Full text] [doi: 10.2196/46694] [Medline: 37163336]
12. OAuth 2.0. URL: https://oauth.net/2/ [accessed 2023-10-27]
13. Zeng K, Bodenreider O, Kilbourne J, Nelson S. RxNav: a web service for standard drug information. AMIA Annu Symp Proc 2006;2006:1156 [FREE Full text] [Medline: 17238775]
14. APIs. National Library of Medicine. URL: https://lhncbc.nlm.nih.gov/RxNav/APIs/index.html [accessed 2023-10-27]
15. Branas R. AngularJS Essentials. Birmingham, UK: Packt Publishing; 2014.
16. Mandel JC, Kreda DA, Mandl KD, Kohane IS, Ramoni RB. SMART on FHIR: a standards-based, interoperable apps platform for electronic health records. J Am Med Inform Assoc 2016 Feb 17:899-908 [FREE Full text] [doi: 10.1093/jamia/ocv189] [Medline: 26911829]
17. Promote interoperability. HealthIT. 2019. URL: https://www.healthit.gov/topic/meaningful-use-and-macra/promoting-interoperability [accessed 2023-10-27]
18. FHIR fact sheets. HealthIT. URL: https://www.healthit.gov/topic/standards-technology/standards/fhir-fact-sheets [accessed 2023-10-27]
19. Kouri A, Yamada J, Lam Shin Cheung J, Van de Velde S, Gupta S. Do providers use computerized clinical decision support systems? A systematic review and meta-regression of clinical decision support uptake. Implement Sci 2022 Mar 10;17(1):21 [FREE Full text] [doi: 10.1186/s13012-022-01199-3] [Medline: 35272667]
20. Gryczynski J, Mitchell SG, Asche SE, Truitt AR, Worley DC, Rindal DB. De-Implementing Opioids for Dental Extractions (DIODE): a multi-clinic, cluster-randomized trial of clinical decision support strategies in dentistry. Implement Sci 2023 Mar 10;18(1):5 [FREE Full text] [doi: 10.1186/s13012-023-01262-7] [Medline: 36765414]
21. HealthPartners. About HealthPartners. URL: https://tinyurl.com/2453fvh8 [accessed 2023-10-26]
22. Rindal DB, Gryczynski J, Asche SE, Truitt AR, Kane SM, Worley DC, et al. De-implementing opioid prescribing in a dental group practice: Lessons learned. Community Dent Oral Epidemiol 2023 Mar;51(1):139-142. [doi: 10.1111/cdoe.12820] [Medline: 36753410]
23. Vorisek CN, Lehne M, Klopfenstein SAI, Mayer PJ, Bartschke A, Haese T, et al. Fast Healthcare Interoperability Resources (FHIR) for interoperability in health research: systematic review. JMIR Med Inform 2022 Jul 19;10(7):e35724 [FREE Full text] [doi: 10.2196/35724] [Medline: 35852842]
24. Sinaci AA, Gencturk M, Teoman HA, Laleci Erturkmen GB, Alvarez-Romero C, Martinez-Garcia A, et al. A data transformation methodology to create findable, accessible, interoperable, and reusable health data: software design, development, and evaluation study. J Med Internet Res 2023 Mar 08;25:e42822 [FREE Full text] [doi: 10.2196/42822] [Medline: 36884270]

Abbreviations

API: application programming interface
CDS: clinical decision support
DIODE: De-Implementing Opioid Use and Implementing Optimal Pain Management Following Dental Extractions
EHR: electronic health record
FHIR: Fast Healthcare Interoperability Resources
HL7: Health Level 7
MED-RT: Medication Reference Terminology
NIH: National Institutes of Health
SMART: Substitutable Medical Applications and Reusable Technologies

Edited by C Perrin; submitted 12.01.23; peer-reviewed by S Delaigue, A Kouri, J Cimino; comments to author 07.03.23; revised version received 24.04.23; accepted 18.10.23; published 07.11.23.

Please cite as:
Rindal DB, Pasumarthi DP, Thirumalai V, Truitt AR, Asche SE, Worley DC, Kane SM, Gryczynski J, Mitchell SG
Clinical Decision Support to Reduce Opioid Prescriptions for Dental Extractions using SMART on FHIR: Implementation Report
JMIR Med Inform 2023;11:e45636
URL: https://medinform.jmir.org/2023/1/e45636
doi:10.2196/45636
PMID:37934572

Review

Scalable Causal Structure Learning: Scoping Review of Traditional and Deep Learning Algorithms and New Opportunities in Biomedicine

Pulakesh Upadhyaya[1,2], PhD; Kai Zhang[1], PhD; Can Li[1], MSPH; Xiaoqian Jiang[1], PhD; Yejin Kim[1], PhD

[1]School of Biomedical Informatics, University of Texas Health Science Center at Houston, HOUSTON, TX, United States

[2]Department of Biomedical Informatics, Emory University School of Medicine, Atlanta, GA, United States

Corresponding Author:
Pulakesh Upadhyaya, PhD
Department of Biomedical Informatics
Emory University School of Medicine
101 Woodruff Circle
Suite 4127
Atlanta, GA, 30322
United States
Phone: 1 9794225161
Email: pulakeshupadhyaya@gmail.com

Abstract

Background: Causal structure learning refers to a process of identifying causal structures from observational data, and it can have multiple applications in biomedicine and health care.

Objective: This paper provides a practical review and tutorial on scalable causal structure learning models with examples of real-world data to help health care audiences understand and apply them.

Methods: We reviewed traditional (combinatorial and score-based) methods for causal structure discovery and machine learning–based schemes. Various traditional approaches have been studied to tackle this problem, the most important among these being the *Peter* Spirtes and *Clark* Glymour algorithms. This was followed by analyzing the literature on score-based methods, which are computationally faster. Owing to the continuous constraint on acyclicity, there are new deep learning approaches to the problem in addition to traditional and score-based methods. Such methods can also offer scalability, particularly when there is a large amount of data involving multiple variables. Using our own evaluation metrics and experiments on linear, nonlinear, and benchmark Sachs data, we aimed to highlight the various advantages and disadvantages associated with these methods for the health care community. We also highlighted recent developments in biomedicine where causal structure learning can be applied to discover structures such as gene networks, brain connectivity networks, and those in cancer epidemiology.

Results: We also compared the performance of traditional and machine learning–based algorithms for causal discovery over some benchmark data sets. Directed Acyclic Graph-Graph Neural Network has the lowest structural hamming distance (19) and false positive rate (0.13) based on the Sachs data set, whereas Greedy Equivalence Search and Max-Min Hill Climbing have the best false discovery rate (0.68) and true positive rate (0.56), respectively.

Conclusions: Machine learning–based approaches, including deep learning, have many advantages over traditional approaches, such as scalability, including a greater number of variables, and potentially being applied in a wide range of biomedical applications, such as genetics, if sufficient data are available. Furthermore, these models are more flexible than traditional models and are poised to positively affect many applications in the future.

(JMIR Med Inform 2023;11:e38266) doi:10.2196/38266

KEYWORDS

causal inference; causal structure discovery; deep learning; biomedicine; networks

Introduction

Background

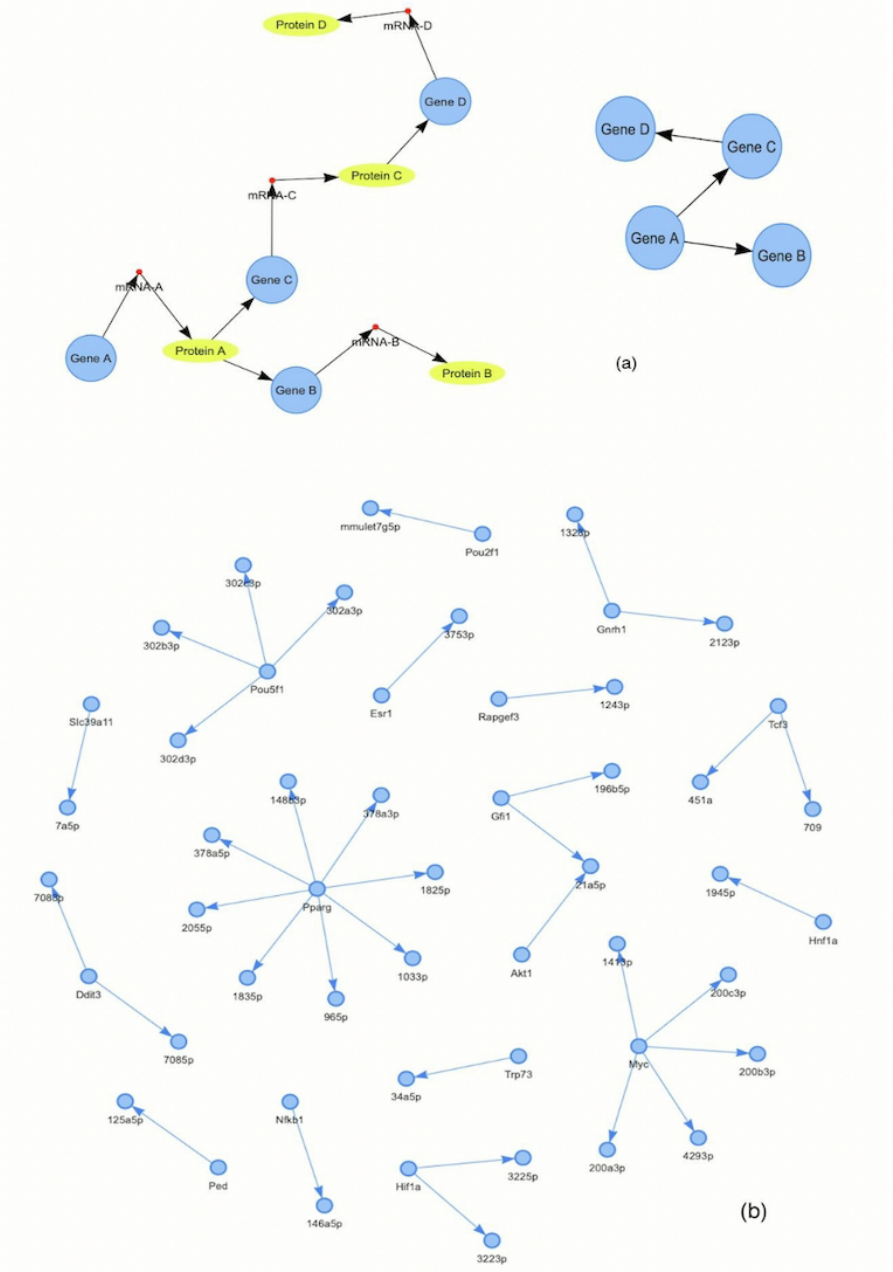
Many applications in biomedicine require the knowledge of the underlying causal relationship between various factors beyond association or correlation. Randomized controlled trials are widely used to uncover causality, but these experiments can be prohibitively expensive or unethical in many cases. Therefore, it has sparked an enormous amount of interest in identifying causal effects from observational data [1-3].

In this paper, we discuss causal structure learning; that is, learning causal relationships that are represented as directed graph structures between different factors and its application to biomedicine. The causal structure is represented by a causal graph (also called a causal Bayesian network), which is a directed acyclic graph (DAG), in which the nodes represent variables and edges represent causation (Figure 1). An edge is drawn from a variable that represents the cause to a variable that represents the effect of that cause. Based on a variety of methodologies, causal structure learning identifies which causal models represented by DAGs accurately represent the observed data.

For example, consider the example of a gene regulatory network [4-7], which is an abstract representation of the gene regulation processes as shown in Figure 1. By observing the data of multiple variables such as gene expression profiles, causal structure learning attempts to discover causal relationships among the genes. For example, if a gene A regulates another gene B, it is represented by an arrow between gene A and gene B.

Many researchers in the biomedical field are interested in causality and not just correlation (eg, whether a particular treatment affects a particular outcome). Unlike association- or correlation-based studies that simply indicate that any 2 variables are correlated, this approach seeks to determine the directional relationship between any 2 variables (eg, between a treatment variable and an outcome variable). In biomedicine, causal structure learning can be applied in a variety of applications.

Figure 1. Example of the causal structure. (A) A gene regulatory network is an abstracted structure (given by the directed graph on the right) of the complex biophysical process shown on the left. (B) A gene regulatory structure from the transmiR database for mice [6].



Examples

Gene Regulatory Networks

A gene regulatory network is a network in which molecular regulators and genes are the nodes, and the directed edges denote the interactions among them [5]. This is in contrast to association-based methods such as finding correlation or mutual information among the genes (finding Pearson, Kendall, Spearman correlation coefficients, etc) that do not have any directional information [8]. Such methods can only be accurate to a certain extent when it comes to deducing extensive gene regulatory structures from data sets with a large set of observations. Correlational studies can only indicate gene-gene association and not the direction of regulation. A gene regulatory network is an example of a causal structure that can be used to develop interventions to control gene expression.

Causal structure learning algorithms have been used to jointly deduce the phenotype network structure and directional genetic architecture [9]. It uses a difference causal inference method and compares it with another causal structure learning algorithm (difference-based Greedy Equivalence Search [GES]) as a baseline. Another study proposed a hybrid algorithm that combines Simulated Annealing with Greedy Algorithm to predict intergene transcriptional regulatory relationships [10], which are also directional in nature. In cancer, somatic genome alterations and differentially expressed genes have causal relationships. A correlational study cannot provide directional information in any of these applications.

The tumor-specific causal inference algorithm proposed by Xue et al [11] uses a Bayesian causal learning framework to find those relationships. Unlike association-based studies, this study is based on a causal structure learning framework across the

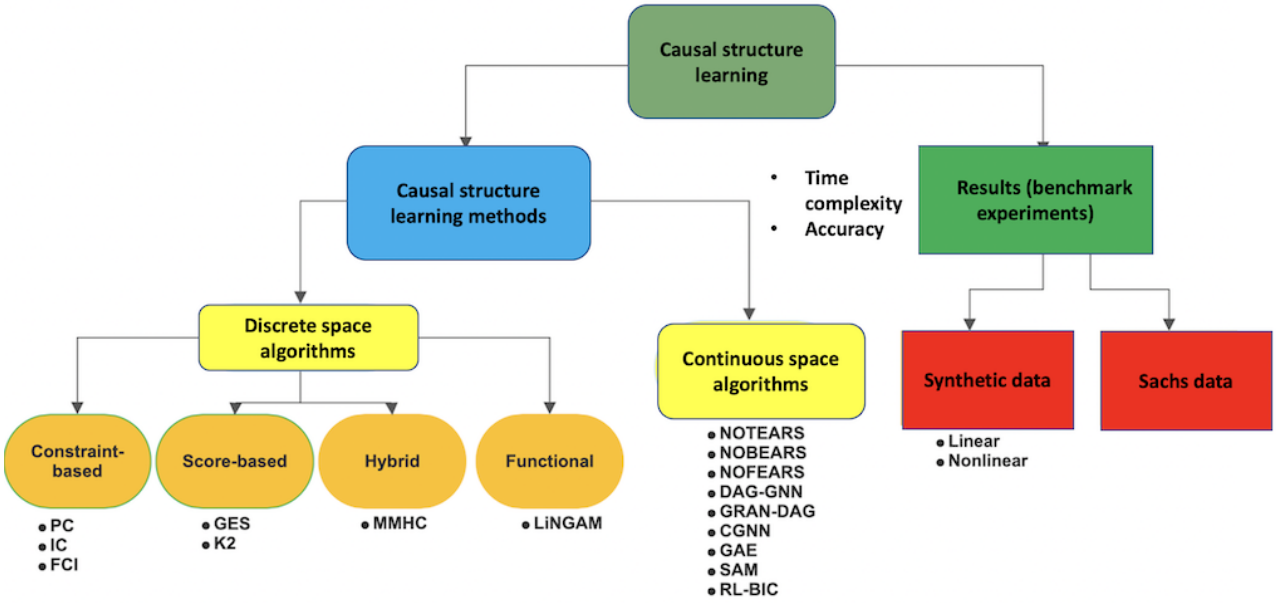
whole genome where Ha et al [12] found gene signatures that were the causes of clinical outcomes and were not merely correlated to them. Apart from these examples, there are also networks such as those represented in the Sachs data set [13] that simultaneously incorporates measurements of 11 phosphorylated proteins and phospholipids to find causal pathways linking them. This is different from association-based correlation studies because protein signaling pathways are directional.

In our comparative analysis of the performance of this data set, we found that machine learning models can also be effective at finding causal structures (details are available in the *Results* section). In the case of more complicated protein signaling networks with many nodes, machine learning–based methods might be particularly effective.

Brain Connectivity Networks

Different regions of the brain have distinct functions. Previous studies have used correlation-based methods [14] to find nondirectional functional connectivity among cortical regions. Spatial localization of brain functions has been studied using methods such as functional magnetic resonance imaging [15]. Regions within the brain are the nodes, and a directed edge between regions represents some functional connection (see Figure 2 in the paper by Brovelli et al [16] for the difference between coherence and causality graphs). Such connections are directional, can have different strengths (weights), and can be both inhibitory or excitatory [17]. Scalable causal structure learning models can also model such connection strengths in addition to directionality, which makes them more expressive than an association. In addition, brains have large-scale structural cortical networks that are directional with respect to information flow and can only be captured by causal structure instead of correlation.

Figure 2. Overview of the methods reviewed and benchmark results sections of the paper. CGNN: Causal Generative Neural Networks; DAG-GNN: Directed Acyclic Graph-Graph Neural Network; FCI: Fast Causal Interface; GAE: graph autoencoder; GES: Greedy Equivalence Search; GRAN-DAG: Gradient-based neural-directed acyclic graph learning; IC: inductive causation; LiNGAM: linear non-Gaussian acyclic model; MMHC: Max-Min Hill Climbing; PC: Peter Spirtes and Clark Glymour; RL-BIC: Reinforcement Learning-Bayesian Information Criterion; SAM: Structural Agnostic Modeling.



Epidemiology

Causal structure learning has also been used in epidemiology with patients' medical records. Many complex diseases are multifactorial in which a combination of these factors contributes to disease predisposition. Causal structure learning considers multiple confounders to determine causal effects solely from one factor of interest to another. For example, causal structure has been used to disentangle psychological factors that predispose adolescents to smartphone addiction [18]. Incorporating a large set of medical claim records, a recent study used a scalable causal structure learning to elucidate the clinical pathways from comorbid illnesses to Alzheimer disease [19].

Challenges

However, there are a few challenges. The general approach to solving this problem of learning a DAG from data, which has been studied for a long time [20], has a time complexity that scales exponentially with the number of observed variables. This is because the problem is generally nondeterministic polynomial-time complete [21]. In practice, if the number of variables is greater than a few hundred, the problem becomes intractable to solve optimally.

Several approaches have been used to solve this problem of intractable time complexity. Traditionally, constraint-based and score-based methods, which search for the optimal graph from a discrete space of candidate graphs, have been used to learn the DAG from data. Constraint-based methods such as the *Peter* Spirtes and *Clark* Glymour (PC) and Fast Causal Inference

(FCI) algorithms (which will be discussed in detail in the *Discrete Space Algorithms* section) rely on statistical tests to estimate the correct causal structure. However, biological data usually involve hundreds to thousands of variables, and the complexity of algorithms increases exponentially as the number of variables increases. For example, typical human RNA sequence data contain at least 20,000 genes. Therefore, the complexity of the PC algorithm is proportional to 2^{20000}, which is infeasible within a reasonable amount of time.

Hence, researchers have investigated various score-based methods that assign scores based on the data to each DAG and select the one with the best score. Although score-based methods scale better than constraint-based methods, they do not scale well for several thousand variables. On the other hand, patient medical records in electronic health records or claim data raise severe scalability concerns, because they include up to 144,000 International Classification of Diseases-Ninth Revision or 69,823 International Classification of Diseases-Tenth Revision diagnosis codes, >2000 US Food and Drug Administration–approved drugs, and >10,000 Current Procedural Terminology procedures or laboratory test codes.

To overcome the limited scalability of traditional methods, recent advances in machine learning algorithms have relaxed the problem of finding an optimal DAG into a continuous optimization problem with smooth acyclicity constraints. This enables the use of nonheuristic machine learning (including deep learning) algorithms to determine the optimal causal structure. This is a promising development in the field of biomedicine. In this study, we focus on scalable algorithms. Table 1 summarizes the algorithms discussed in this study. The tools available for some of these algorithms are listed in Multimedia Appendix 1. A list of ground truth causal structures can be found in the *bnlearn* repository [22].

There are 2 distinct approaches in the context of treatment effect evaluation: the structural approach and potential outcome framework approach [23]. In this study, we consider the first approach, in which there are 2 distinct types of algorithms for finding the causal DAG structure. In all of these examples, the goal is to learn a DAG that shows the directional relationship among variables from observational data.

Table 1. Summary of various algorithms for causal structure learning.

Algorithm	DS[a]	CS[b]	Summary	Remarks	Scalability
PC[c]	✓	✗	A partially directed acyclic graph (CPDAG[d]) is produced by iteratively checking the conditional independence conditions of adjacent nodes, conditioned on an all-size subset of neighbors.	Outputs usually converge to the same equivalence class; high FPR[e] on experimental data	+[f]
IC[g]	✓	✗	Returns the equivalent class of the DAG[h] based on the estimated probability distribution of random variables and an underlying DAG structure.	Outputs usually converge to the same equivalence class.	+
FCI[i]	✓	✗	Modified PC algorithm to detect unknown confounding variables and produces asymptotically correct results.	Faster than PC with similar TPR[j]; converges to the same asymptotic result; high experimental FPR	++
GES[k]	✓	✗	Starts with an empty graph and iteratively adds and deletes edges in the graph by optimizing a score function.	Faster than PC with higher TPR; stable result for the same score function	++
Fast GES	✓	✗	Improved and parallelized version of GES	Faster than GES; same TPR; stable result for the same score function	++
K2	✓	✗	Performs a greedy heuristic search for each nodes' parents.	Greedy searches might return very suboptimal solutions.	++
MMHC[l]	✓	✗	MMHC to find the skeleton of the network and constrained greedy search for edge orientation.	Greedy searches might return suboptimal solutions.	+
LiNGAM[m]	✓	✗	Transfer the linear structure model  to the form of  , and optimize for matrix B.	Works very well on linear data but not on nonlinear data.	++
NOTEARS	✗	✓	Uses smooth function $h(A)$, whose value characterizes the "DAG-ness" of the graph with adjacency matrix A—that is, $h(A)=0$ for DAG—and optimizes using continuous optimization.	Might converge to many different DAGs; GPUs[n] can speed up the process.	+++
NOBEARS	✗	✓	Proposed a new acyclicity constraint that allows for faster optimization and scalability, and a polynomial regression loss to infer gene regulatory networks from nonlinear gene expressions.	Might converge to many different DAGs; GPUs can speed up the process.	+++
DAG-GNN[o]	✗	✓	Uses an autoencoder framework and deep learning to train it and infer the causal structure from the weights of the trained network and is more scalable than NOTEARS.	Might converge to many different DAGs; GPUs can speed up the process.	++++
NOFEARS	✗	✓	Modify NOTEARS so the scoring function remains convex to ensure local minima.	Might converge to many different DAGs; GPUs can speed up the process.	++++
GAE[p]	✗	✓	Scalable graph autoencoder framework (GAE) whose training time increases linearly with the number of variable nodes.	Good accuracy; might converge to many different DAGs; GPUs can speed up the process.	++++
GRAN-DAG[q]	✗	✓	Extends the NOTEARS algorithm for nonlinear relationships.	Works on nonlinear data; better accuracy than NOTEARS; might converge to many different DAGs; GPUs can speed up the process.	++++
CGNN[r]	✗	✓	Generative model of the joint distribution of variables reducing MMD[s] between the graph and data.	Does not always converge to a single class of equivalent DAGs; GPUs can speed up the process.	++++
SAM[t]	✗	✓	Structurally agnostic model for causal discovery and penalized adversarial learning.	Does not always converge to a single class of equivalent DAGs; GPUs can speed up the process.	++++
RL-BIC[u]	✗	✓	Reinforcement learning-based algorithm that uses both the acyclicity constraint and the BIC[v] score.	Very good accuracy; does not always converge to a single class of equivalent DAGs; GPUs can speed up the process.	++++

[a]DS: discrete space algorithms.
[b]CS: continuous space algorithms.
[c]PC: *Peter* Spirtes and *Clark* Glymour.
[d]CPDAG: completed partially directed acyclic graph.
[e]FPR: false positive rate.
[f]The + symbol for an algorithm indicates its scalability.
[g]IC: inductive causation.
[h]DAG: directed acyclic graph.
[i]FCI: Fast Causal Inference.
[j]TPR: true positive rate.
[k]GES: Greedy Equivalence Search.
[l]MMHC: Max-Min Hill Climbing.
[m]LiNGAM: linear non-Gaussian acyclic model.
[n]GPUs: graphical processing units.
[o]DAG-GNN: Directed Acyclic Graph-Graph Neural Network.
[p]GAE: graph autoencoder.
[q]GRAN-DAG: Gradient-based neural - directed acyclic graph learning.
[r]CGNN: causal generative neural network.
[s]MMD: maximum mean discrepancy.
[t]SAM: Structural Agnostic Modeling.
[u]RL-BIC: Reinforcement Learning-Bayesian Information Criterion.
[v]BIC: Bayesian information criterion.

Paper Structure

This study attempts to provide a comparative study of various scalable algorithms that are used to discover causal structures from observational data to the biomedicine community. Some of these traditional and score-based methods have been extensively studied [24], but many of the algorithms discussed here focus on scalable causal structure learning. Although we do not list all possible approaches as Vowels et al [25], we sample a few important algorithms and evaluate their performance on synthetic data sets and the Sachs data set [13].

This tutorial paper presents algorithms for causal structure identification in biomedical informatics. In the *Methods* section, we discuss the methodology and examine the traditional algorithms that determine the optimal causal graph in a discrete space. We also discuss algorithms that use continuous space optimization to discover causal relationships. We compare the performance of these algorithms in the *Results* section. Finally, we present the discussion and conclusions. A brief overview of the methods and results is presented in Figure 2.

Methods

Overview

In this section, we discuss 2 paradigms of algorithms for causal structure learning. First, we consider algorithms that search for the optimal DAG in the discrete space of all possible DAGs (space of all possible discrete DAGs for a given number of variable nodes) or *discrete space algorithms*. Second, we consider scalable algorithms that use continuous optimization methods to find the optimal DAG (ie, algorithms that search the continuous space of all possible weighted DAGs to find the optimal one), known as *continuous space algorithms*.

Discrete Space Algorithms

Overview

The first type that we discuss in this section is discrete space algorithms for causal discovery; that is, algorithms that search for the optimal DAG in the discrete space of candidate causal graphs. This is in contrast to continuous space algorithms (discussed in the *Continuous Space Algorithms* section) that search for the optimal DAG from the continuous space of weighted candidate graphs.

The discrete space algorithms can be divided into the following 4 types: combinatorial constraint-based models, score-based models, hybrid models, and functional models. In combinatorial constrained-based methods, we consider methods that check the conditional independence relations of 2 adjacent nodes conditioned on all subsets of their neighbors. Such methods can be useful when the number of variables is up to a few hundred. Score-based methods perform optimization by considering a score representing the goodness of fit and can handle more variables than constraint-based methods. Hybrid methods combine constraint- and score-based algorithms. Functional models find structural equations to describe the causal relationship and are useful mostly when the variables can be assumed to be expressed by some linear or nonlinear equations.

Combinatorial Methods

We now focus on combinatorial optimization methods, where conditional independence relationships in the data are used for finding the optimal DAG.

PC and Its Variants

The PC algorithm was proposed by *PC* and is named after them [26]. This algorithm produces a completed partially DAG (CPDAG) by iteratively checking the conditional independence relations of 2 adjacent nodes conditioned on all-size subsets of

their neighbors. Three assumptions underlie the algorithm: no confounder variable, the causal Markov condition, and faithfulness. Under these conditions, this algorithm generates a partially directed causal graph that is proven to be asymptotically correct.

The PC algorithm is order-dependent; that is, the output of the algorithm can depend on the order in which the variables are provided to the algorithm. To address this problem, Colombo and Maathuis [27] developed a PC-stable algorithm in which the deletion of an edge takes place at the end of each stage (considering any 2 nodes' relations within a predetermined neighborhood). Thus, any ordering of vertices will result in the same edge deletions, resulting in the same stable output. The PCMCI and PCMCI+ [27-29] are 2 extensions of the PC algorithm proposed to handle large-scale time-series data sets.

Inductive Causation Algorithm and Its Variants

The inductive causation (IC) algorithm uses the estimated probability distribution of random variables with an underlying DAG structure and outputs the equivalent class of the DAG. In contrast, PC provides a schematic search method and is thus considered a refinement of the IC.

The IC* algorithm [30,31] is an extension of the IC algorithm, which searches for causal relations using observations of a set of variables, even when they appear as latent variables. The output of the IC algorithm is a CPDAG that only has directed edges (identified causation) and undirected edges (undetermined causation). The output of the IC* algorithm is an embedded pattern; that is, a hybrid graph containing ≥2 types of edges.

FCI and Its Variants

The FCI is a modification of the PC algorithm [30,32] that detects unknown confounding variables and produces asymptotically correct results. FCI improves the PC algorithm by adopting 2 rounds of phases of the PC algorithm. The algorithm first uses PC-phase I to find an initial skeleton, then uses the separation set to orient all v-structure triples (a->c<-b) and outputs a CPDAG; then performs another round of skeleton searching based on the CPDAG and repeats the orientation for unshielded triples. The really fast causal inference algorithm [33,34] skips the second step, which is the most time-consuming part of the task, and therefore significantly accelerates the FCI procedure. A set of 10 rules was added to the algorithm to orient the edges of the skeleton.

Score-Based Methods

Overview

In addition to traditional combinatorial methods such as PC and FCI, score-based methods have also been used to uncover causal structures. In these methods, algorithms determine the optimal DAG by optimizing a particular score.

A typical score function is the Bayesian information criterion (BIC) score. The GES algorithm uses different score functions for different data types as follows: the BIC score (for continuous data), likelihood-equivalence Bayesian Dirichlet uniform joint distribution score (for discrete data), and Conditional Gaussian score (for continuous or discrete mixture data).

where  is the maximized likelihood function of the model, n is the number of observational data points, and k is the degree of freedom. The definition of the Bayesian Dirichlet uniform joint distribution scoring function was found in a study by Buntine [35]. The conditional Gaussian score is defined on the ratios of joint distributions, and Andrews et al [36] have proved that the Conditional Gaussian score is *score equivalent*; that is, a scoring function that scores all DAGs in the same Markov Equivalence Class equally.

Score-based methods include the GES algorithm, the fast GES algorithm, and the K2 algorithm.

GES Algorithm

The GES algorithm was proposed by Chickering [37], and its underlying principles were obtained from Meek [37,38]. The algorithm starts with an empty graph and iteratively adds and deletes the edges in the graph by optimizing a score function. During the forward phase, the algorithm searches iteratively from the space of the DAGs created by one edge addition on the current DAG and selects the edge with the best score. The forward phase ends when the score is no longer increasing. In the second phase, the algorithm repeats the above step but deletes one edge at a time and selects the edge that improves the score the most. The algorithm stops as soon as there are no more edges to be deleted.

Fast GES Algorithm

Fast GES is an improved and parallelized version of the GES. Significant speedup was achieved by storing the score information during the GES algorithm [39]. In addition, several insights regarding parallelization were offered in the paper. First, the precalculation of covariances can be parallelized by variables. Second, it is possible to parallelize the process of calculating the edge scores when an edge addition is being performed on the graph. A greater speedup can be achieved for sparse graphs.

K2 Algorithm

The main idea of the K2 algorithm [40] is to perform a greedy heuristic search of the parents of each node. For each node, the algorithm iteratively determines the parents. When visiting node X_i, the algorithm searches for all possible parents of X_i (X_j such that j has a lower ordering of i). The algorithm greedily adds X_j to the parent set of X_i if it could increase a predefined score function. The iteration for node X_i stops when the number of parent nodes reaches the (preset) maximum or when adding an X_j does not increase the score anymore. The entire algorithm finishes after completing the iteration for all X_i.

Hybrid Algorithms

Hybrid algorithms use a combination of score-based and combinatorial constraint-based optimization methods to determine the optimal DAG. An example is the Max-Min Hill Climbing (MMHC) algorithm**.** The MMHC algorithm is a combination of constraint- and score-based algorithms [41]. It uses the Max-Min Parents and Children algorithm. A detailed

description is provided in [41,42] to find the skeleton of the Bayesian network and then perform the constrained greedy search to orient the edges.

Algorithms for Functional Causal Models

Overview

Functional causal models or structural equation models (SEMs) assume structural equations that define the causal relationships. Such structural equations may describe the linear and nonlinear relationships among variables. In addition to the discrete methods discussed here, SEMs are also an important assumption in many machine learning–based methods that use the continuous optimization techniques in the *Continuous Space Algorithms* section.

Linear Non-Gaussian Acyclic Model

The linear non-Gaussian acyclic model (LiNGAM) was originally proposed by Shimizu [43] to learn linear non-Gaussian acyclic causal graphs from continuous-valued data. The LiNGAM transfers  to the form of  , and the causal structure problem becomes an optimization problem for matrix *B*. There are several extensions of the LiNGAM model using different estimation methods, including independent component analysis–based LiNGAM [43,44], DirectLiNGAM [45], and Pairwise LiNGAM [46].

Additive Noise Models

A nonlinear additive noise model is proposed in [47]. The model assumes that the observed data are generated according to the following equation:

where f_i is an arbitrary function, $x_{pa(i)}$ denotes the ancestor nodes of node x_i in the true causal graph, and n_i is the noise variable of an arbitrary probability density function. This study proves the basic identifiability principle for the 2 variables case and generalizes the results to multiple variables.

Continuous Space Algorithms

Overview

Traditional causal discovery algorithms attempt to discover a causal graph, which is usually a DAG, while searching for an optimal graph in the space of candidate graphs. The score-based optimization problem of DAG learning (discussed in the *Score-Based Methods* section) is mathematically given by the following equation:

Here,  is the set of all DAGs with *d* nodes, and  is the cost or score function. The problem of searching for all DAGs is usually intractable and superexponential in the number of nodes in the graph.

An alternative approach would be to model the problem as a continuous space optimization problem, which would then allow the application of various learning techniques. Recently, several publications have explored continuous optimization methods that learn DAGs by adding an acyclicity constraint. In these approaches, the discrete acyclicity constraint  is replaced by  , where $h(A)$ is a smooth function that ensures acyclicity of $G(A)$.

The hard constraints on acyclicity can be relaxed and incorporated into the loss function to be optimized. This smooth continuous constraint allows the use of machine learning–based tools, which in turn can make the algorithms scalable in the presence of substantial amounts of data. These algorithms are based on SEMs.

NOBEARS Algorithm

Several other improvements such as the NOBEARS algorithm [48] have improved the scalability of the NOTEARS algorithm. A fast approximation of a new constraint is proposed, and a polynomial regression loss model is proposed to account for nonlinearity in gene expression to infer gene regulatory networks.

NOTEARS Algorithm

This algorithm considers the acyclicity constraint and comes up with the constraint.

Here,  is the element-wise product. $h(A)$ equals 0 if and only if $G(A)$ is acyclic, and more severe deviations from acyclicity would increase the value of the function. This study assumes a linear SEM:

Here, X_i is a d-dimensional sample vector of the joint distribution of d variables and Z_i is a d-dimensional noise vector. We denote n such samples by matrix X, and the loss function (with l_1-regularization) is given as follows:

The constraint is given by $h(A)=0$ and is used in the final Lagrangian formulation of the loss function. The paper on learning sparse nonparametric DAGs is an extension of NOTEARS, which tries to define a “surrogate” of the matrix above for general nonparametric models to optimize [49].

Directed Acyclic Graph-Graph Neural Network Algorithm

A Directed Acyclic Graph-Graph Neural Network (DAG-GNN) [50] generalizes the NOTEARS algorithm by considering the nonlinearity in the SEMs. It can be modeled with a variational autoencoder neural network with a special structure, with an encoder  , and a decoder  and where g_1,g_2 are parameterized functions that can be assumed to serve as the inverse of f_1,f_2, respectively.

This variational framework considers Z to be a latent vector (instead of viewing it as noise in linear SEMs), which can have dimensions other than d. The decoder then attempts to

reconstruct the data from this latent variable. The encoder and decoder can be trained together from *n* samples of  such that the loss function:

is minimized, where *KLD* is the Kullback-Liebler Divergence. The constraint in this optimization process to ensure the acyclicity of matrix *A* is slightly modified to:

where α is an arbitrary parameter. This constraint can be implemented more easily in graphical processing unit-based deep learning libraries owing to the algorithm's parallelizability and scalability of the algorithm.

NOFEARS Algorithm

Wei et al [51] demonstrated that the NOTEARS algorithm fails to satisfy the Karush-Kuhn-Tucker regularity conditions. Therefore, they reformulated the problem to ensure that the convexity of the scoring function can still ensure local minima even when the constraints are nonconvex. This new algorithm called the NOFEARS algorithm has the following acyclicity constraint.

Graph Autoencoder

Ng et al [52] propose another graph autoencoder (GAE) framework for causal structure learning, which improves the training speed and performance over DAG-GNN for both linear and nonlinear synthetic data sets.

Some other similar machine learning–based continuous learning algorithms include gradient-based neural DAG [53], Causal Generative Neural Network [54], and structurally agnostic model [55].

Reinforcement Learning-Based Methods

Reinforcement learning-based methods have been proposed recently that consider both the acyclicity constraint and BIC score in the reward function and attempt to learn the DAG [56]. They used an actor-critic model, where the actor is an encoder-decoder framework that takes data as input and outputs the graph. The critic uses the reward function for this graph and updates the proposed graph.

Results

This section provides the results to compare the effectiveness of some causal structure learning algorithms on synthetic and real data.

Benchmark Methods

The synthetic data were generated in the same manner as in the DAG-GNN paper [50]. An Erdos-Renyi model with an expected node degree of 3 was used to generate the random graph, and the adjacency matrix was formed by assigning weights to the edges from a uniform distribution. The samples were generated using the following structural equation:

Here, *Z* is random Gaussian noise. We consider 2 functions for *g*(*X*). The first is a (linear) identity function:

and the second is a nonlinear function

We considered 5 data sets for both linear and nonlinear functions. For each data set, we generated n=5000 independent samples according to the above equations. We used 6 algorithms, 4 of which are discrete space algorithms. PC and Greedy Fast Causal Interface (GFCI) are constraint-based methods, GES is a score-based method, and MMHC is a hybrid method. We also considered 2 continuous space methods, DAG-GNN and GAE.

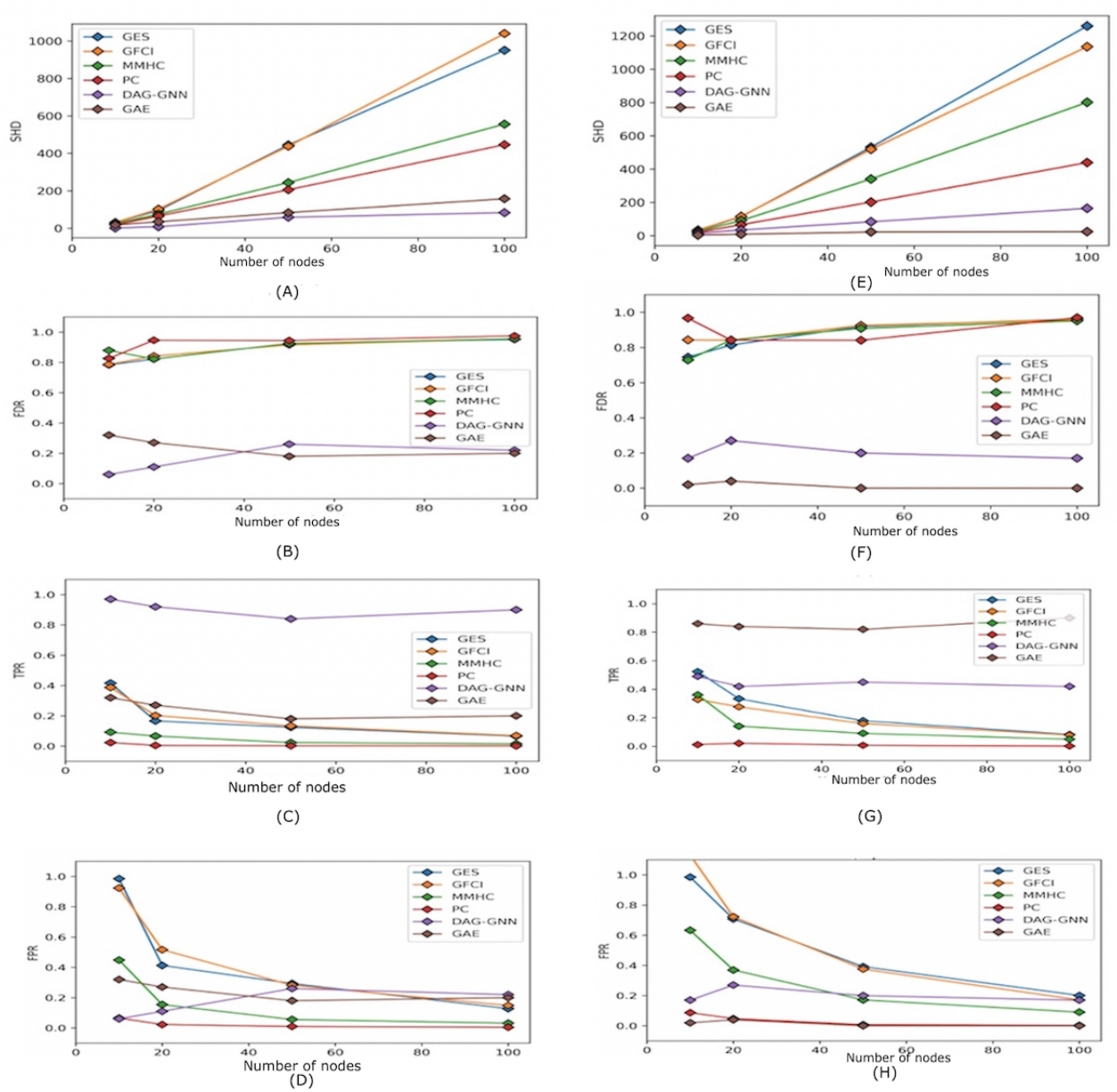
We also evaluated these algorithms on the publicly available Sachs data set [13] using the above 4 metrics and showed the results are shown in Table 2. For other data sets that have the ground truth but are not covered in our experiments, please refer to the *bnlearn* repository [22]. The algorithms were implemented in Python for machine learning–based continuous space methods and R for discrete space algorithms.

Table 2. Benchmark experiments on the Sachs data set. We evaluated 6 algorithms— Peter Spirtes and Clark Glymour (PC), Greedy Equivalence Search (GES), Greedy Fast Causal Interface (GFCI), Max-Min Hill Climbing (MMHC), Directed Acrylic Graph-Graph Neural Network (DAG-GNN), and graph auto encoder (GAE), on 4 metrics—structural hamming distance (SHD), true positive rate (TPR), false positive rate (FPR), and false discovery rate (FDR)—and show their results in Figure 3. In all these evaluations, we consider any edge whose direction is reversed as half discovered.

Metric	PC	GES	GFCI	MMHC	DAG-GNN	GAE
SHD (↓)	24.50	26.50	29.50	22.00	*19.00* [a]	22.00
FDR (↓)	0.77	0.72	0.79	*0.68*	0.71	0.89
TPR (↑)	0.32	*0.56*	0.44	0.47	0.11	0.05
FPR (↓)	0.49	0.64	0.72	0.45	*0.13*	0.21

[a]Italicized values represent the best results for each metric.

Figure 3. Accuracy comparison. We evaluated 6 algorithms: Peter Spirtes and Clark Glymour (PC), Greedy Equivalence Search (GES), Greedy Fast Causal Interface (GFCI), Max-Min Hill Climbing (MMHC), Directed Acrylic Graph-Graph Neural Network (DAG-GNN), and graph auto encoder (GAE) on 4 metrics, structural hamming distance (SHD↓), true positive rate (TPR↑), false positive rate (FPR↓), and false discovery rate (FDR↓). In all these evaluations, we considered any bidirectional edges as half discovered. In experiments (A-D), first column, the data are drawn from a distribution according to the underlying causal graph where relationships between nodes are linear, and experiments (E-H), second column, is for the nonlinear case. In all experiments, the number of nodes of the graph ranges from 10, 20, 50, to 100. For each graph size, we drew 5 different data sets from the graph structure with a sample size of 1000 and calculated 4 evaluation metrics and obtained the average.



Observations

All algorithms were tested on both linear and nonlinear data. The accuracies of some of these algorithms are shown in Figure 3. We used the following 4 evaluation measures: structural hamming distance (SHD), true positive rate (TPR), false positive rate (FPR), and false discovery rate (FDR). SHD refers to the number of edge insertions, deletions, and reversals. In our case, we used a modified SHD, where a reversal contributes half of the SHD score instead of 1. The TPR is the ratio of the algorithm's correctly discovered edges to the number of edges in the ground truth graph. FPR is the ratio of the algorithm's falsely discovered edges to the number of nonedges in the ground truth graph. FDR is the ratio of the algorithm's falsely discovered edges to the total number of discovered edges. We also evaluated these algorithms on the Sachs data set using the above 4 metrics, and the results are shown in Table 2.

Time Complexity

The relative scalability of different algorithms is presented in Table 1. The worst-case time complexity of the PC, GES, and GFCI algorithms was $O(m^k)$, where m is the number of variables (nodes in the DAG). For the GES, the best-case time complexity was $O(m)$. For the GAE and DAG-GNN, the time complexity of the algorithm is $O(m^2+k)$, where k is the number of iterations. The time complexity for MMHC is $O(mV)$, where V is the set of

variables, *S* is the largest set of parents and children, and *l* is a parameter of the algorithm that denotes the size of the largest conditioned subset [41].

In our experiments, we observed that in the worst-case scenario, the running time for a maximum of 100 variables was of the order of hours for MMHC and of the order of minutes for the other algorithms. However, as the number of variables increases to a few thousand, machine learning–based methods such as DAG-GNN and GAE can provide solutions in a reasonable time. The trade-off between complexity (number of iterations) and accuracy can provide a choice between a method that is less accurate but faster or vice versa.

Discussion

Interpretation of Results

It is clear from the results that the algorithms have different advantages and disadvantages. Although the PC algorithm performs well across both linear and nonlinear data, it has a low TPR and is computationally intensive. The GES, GFCI, and MMHC algorithms show a very high FPR, but their TPR is higher than that of the PC algorithm. The SHD of the 2 machine learning–based methods—DAG-GNN and GAE—was also considerably lower for both data sets.

Continuous constraint-based algorithms generally exhibit a very low FDR, except for the benchmark Sachs data set. This is generally because both linear and nonlinear models are based on SEMs with the same causal relationship function at every node, which is an algorithm assumption when they learn the causal structure, but one cannot guarantee the same for Sachs data [13], because such constraints cannot be defined a priori.

This is corroborated by recent results from Zhu et al [56] where such gradient-based methods performed poorly on data generated by a nonlinear model, in which every causal relationship (node function) was sampled from a Gaussian distribution. However, this is a growing research area. In general, in areas such as gene regulatory networks and brain connectivity networks where the number of variables is large, machine learning–based methods can provide comparable results to traditional methods with a much more efficient time complexity and scalability.

Challenges

Machine learning for causal structure learning is not without its limitations, which may present several challenges. First, in many applications, there is no ground truth about causal structure, which makes it difficult to evaluate the performance of these algorithms. Furthermore, many scalable methods use stochastic gradient descent; thus, the final output graph is not always deterministic. When the number of data samples or variables is low, traditional or score-based methods are a better choice, especially when the application requires fewer false positives. For the PC, GES, and GFCI algorithms, we observed that these algorithms require considerable running time, as the number of variables is more than 100 [57].

However, when it comes to large samples of data (eg, more than 100,000 samples) or hundreds of variables (eg, in many gene networks), machine learning methods can provide a reasonable solution, because other methods fail owing to scalability issues. As machine learning algorithms are highly parallelizable, the solutions can be computed much faster, particularly through the use of a graphical processing unit. These algorithms are potentially useful for many applications related to genetics and biomedicine, especially those with an abundance of observational data.

The continuous space machine learning models are more scalable and might be useful in the era of big data. Traditional methods might have complexities that grow exponentially with the number of attributes. Despite the nonconvexity of the optimization proposed by Zheng et al [58], optimization and learning strategies can be used to help find the optimal solution. Several methods have been used to solve this problem using augmented Lagrangian approaches [50,52].

The NOBEARS algorithm reduces the computing complexity of NOTEARS from cubic to quadratic in terms of the number of attributes [48], allowing for smooth implementation in data sets that have more than 4000 attributes. The algorithms are also highly parallelizable, and most of the algorithms use deep learning libraries such as Tensorflow [59] and PyTorch [60].

Machine learning techniques for causal discovery, which use continuous space optimization, are an emerging area of research, which can lead to more efficient causal discovery, particularly in applications where directed graphs are used to specify causal relations more clearly. With sufficient data, machine learning models can be robust to certain discrepancies such as sample bias, missing data, and erroneous measurements. Many of these applications have also focused on weaker concepts of causality such as pairwise directionality during the analysis of gene networks and brain connectivity networks [61,62].

It is noteworthy that machine learning methods are usually black box methods, which might provide lesser insight into the process of derivability of the causal structures. For higher interpretability, an option that has been explored is to develop parallel versions of these algorithms, such as PC [63]. In the future, options such as ensemble learning can be explored for the same.

Some other challenges can be found in finding causal structure from data. In the case of learning causal structure from electronic health record data, they might have several problems, such as missing values or noise in the data, which are very common [64]. If the number of missing values or the amount of noise is significant, the application of causal discovery methods might yield unreliable results.

Furthermore, most causal discovery methods assume that the distribution of data is stationary, which may not be true in certain medical applications [65]. Hence, it is very important to consider the aforementioned problems as well as issues related to selection bias before causal structure learning methods are applied. Glymour et al [24] discuss some general guidelines to avoid such problems in causal structure learning. These generalized learning algorithms are ineffective in many biomedical applications, such as in learning biological or gene networks, because they do not consider specific network

constraints. These constraints can be incorporated into causal structure learning methods for greater efficiency.

Conclusions

In this paper, we have discussed the motivation for causal structure discovery in biomedicine as well as some interesting applications. Two paradigms of causal discovery algorithms have been reviewed. Combinatorial or score-based algorithms are used in the first paradigm for optimizing discrete spaces of candidate causal graphs, whereas machine learning algorithms are used in the second paradigm to solve continuous optimization problems with acyclicity constraints. In addition to listing these methods, we have also included resources that readers can use to find appropriate applications. Furthermore, we tested several algorithms against synthetic benchmark data sets and against the Sachs real-world data set and evaluated their relative performances. We have also discussed their theoretical time complexity. Our discussion of the limitations and challenges of various algorithms is intended to offer readers a guide for choosing an algorithm from among the many available options. Finally, we highlight several challenges associated with finding causal structure from real-world data (eg, missing values, nonstationarity, noise, and sampling bias).

Acknowledgments

XJ is a Cancer Prevention and Research Institute of Texas scholar in cancer research (RR180012) and was supported in part by the Christopher Sarofim Family Professorship, University of Texas Stars award, University of Texas Health Science Center startup, and the National Institutes of Health under award numbers R01AG066749 and U01TR002062.

Authors' Contributions

The survey and experiments on deep learning–based methods and the survey on potential applications were conducted by PU. The survey and experiments on traditional methods were conducted by KZ and CL. XJ and YK conceived the study and provided useful inputs for the potential applications of scalable structure learning.

Conflicts of Interest

None declared.

Multimedia Appendix 1
List of tools for causal discovery.
[DOCX File , 21 KB - medinform_v11i1e38266_app1.docx]

References

1. Spirtes P, Glymour C, Scheines R, Kauffman S, Aimale V, Wimberly F. Constructing Bayesian network models of gene expression networks from microarray data. Carnegie Mellon University. 2000. URL: https://kilthub.cmu.edu/articles/journal_contribution/Constructing_Bayesian_Network_Models_of_Gene_Expression_Networks_from_Microarray_Data/6491291 [accessed 2021-01-15]
2. Park J, Liang M, Alpert JM, Brown RF, Zhong X. The causal relationship between portal usage and self-efficacious health information-seeking behaviors: secondary analysis of the health information national trends survey data. J Med Internet Res 2021 Jan 27;23(1):e17782 [FREE Full text] [doi: 10.2196/17782] [Medline: 33502334]
3. Tian S, Bi M, Bi Y, Che X, Liu Y. A bayesian network analysis of the probabilistic relationships between various obesity phenotypes and cardiovascular disease risk in Chinese adults: Chinese population-based observational study. JMIR Med Inform 2022 Mar 02;10(3):e33026 [FREE Full text] [doi: 10.2196/33026] [Medline: 35234651]
4. TransmiR v2.0 database. The Cui Lab. URL: http://www.cuilab.cn/transmir [accessed 2021-10-14]
5. Belyaeva A, Squires C, Uhler C. DCI: learning causal differences between gene regulatory networks. Bioinformatics 2021 Mar 11;37(18):3067-3069. [doi: 10.1093/bioinformatics/btab167] [Medline: 33704425]
6. Huynh-Thu VA, Sanguinetti G. Gene regulatory network inference: an introductory survey. Methods Mol Biol 2019;1883:1-23. [doi: 10.1007/978-1-4939-8882-2_1] [Medline: 30547394]
7. Liu Z, Wu C, Miao H, Wu H. RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. Database (Oxford) 2015;2015:bav095 [FREE Full text] [doi: 10.1093/database/bav095] [Medline: 26424082]
8. Liu Z. Quantifying gene regulatory relationships with association measures: a comparative study. Front Genet 2017;8:96 [FREE Full text] [doi: 10.3389/fgene.2017.00096] [Medline: 28751908]
9. Neto EC, Keller MP, Attie AD, Yandell BS. Causal graphical models in systems genetics: a unified framework for joint inference of causal network and genetic architecture for correlated phenotypes. Ann Appl Stat 2010 Mar 01;4(1):320-339 [FREE Full text] [doi: 10.1214/09-aoas288] [Medline: 21218138]

10. Adabor ES, Acquaah-Mensah GK, Oduro FT. SAGA: a hybrid search algorithm for Bayesian Network structure learning of transcriptional regulatory networks. J Biomed Inform 2015 Feb;53:27-35 [FREE Full text] [doi: 10.1016/j.jbi.2014.08.010] [Medline: 25181467]
11. Xue Y, Cooper G, Cai C, Lu S, Hu B, Ma X, et al. Tumour-specific causal inference discovers distinct disease mechanisms underlying cancer subtypes. Sci Rep 2019 Sep 13;9(1):13225 [FREE Full text] [doi: 10.1038/s41598-019-48318-7] [Medline: 31519988]
12. Ha MJ, Baladandayuthapani V, Do K. Prognostic gene signature identification using causal structure learning: applications in kidney cancer. Cancer Inform 2015;14(Suppl 1):23-35 [FREE Full text] [doi: 10.4137/CIN.S14873] [Medline: 25861215]
13. Sachs K, Perez O, Pe'er D, Lauffenburger DA, Nolan GP. Causal protein-signaling networks derived from multiparameter single-cell data. Science 2005 Apr 22;308(5721):523-529. [doi: 10.1126/science.1105809] [Medline: 15845847]
14. Achard S, Salvador R, Whitcher B, Suckling J, Bullmore E. A resilient, low-frequency, small-world human brain functional network with highly connected association cortical hubs. J Neurosci 2006 Jan 04;26(1):63-72 [FREE Full text] [doi: 10.1523/JNEUROSCI.3874-05.2006] [Medline: 16399673]
15. Deshpande G, Hu X. Investigating effective brain connectivity from fMRI data: past findings and current issues with reference to Granger causality analysis. Brain Connect 2012;2(5):235-245 [FREE Full text] [doi: 10.1089/brain.2012.0091] [Medline: 23016794]
16. Brovelli A, Ding M, Ledberg A, Chen Y, Nakamura R, Bressler SL. Beta oscillations in a large-scale sensorimotor cortical network: directional influences revealed by Granger causality. Proc Natl Acad Sci U S A 2004 Jun 29;101(26):9849-9854 [FREE Full text] [doi: 10.1073/pnas.0308538101] [Medline: 15210971]
17. Bielczyk NZ, Uithol S, van Mourik T, Anderson P, Glennon JC, Buitelaar JK. Disentangling causal webs in the brain using functional magnetic resonance imaging: a review of current approaches. Netw Neurosci 2019;3(2):237-273 [FREE Full text] [doi: 10.1162/netn_a_00062] [Medline: 30793082]
18. Kim Y, Jeong J, Cho H, Jung D, Kwak M, Rho M, et al. Personality factors predicting smartphone addiction predisposition: behavioral inhibition and activation systems, impulsivity, and self-control. PLoS One 2016;11(8):e0159788 [FREE Full text] [doi: 10.1371/journal.pone.0159788] [Medline: 27533112]
19. Kim Y, Zhang K, Savitz SI, Chen L, Schulz PE, Jiang X. Counterfactual analysis of differential comorbidity risk factors in Alzheimer's disease and related dementias. PLOS Digit Health 2022 Mar 15;1(3):e0000018. [doi: 10.1371/journal.pdig.0000018]
20. Neufeld E. Judea Pearl. Probabilistic reasoning in intelligent systems: networks of plausible inference. Series in representation and reasoning. Morgan Kaufmann, San Mateo1988, xix + 552 pp. J Symb Log 2014 Mar 12;58(2):721. [doi: 10.2307/2275238]
21. Chickering D. Learning Bayesian networks is NP-complete. In: Learning from Data. New York: Springer; 1996.
22. Bayesian network repository. BN Learn. URL: https://www.bnlearn.com/bnrepository/ [accessed 2022-03-24]
23. Ling Y, Upadhyaya P, Chen L, Jiang X, Kim Y. Heterogeneous treatment effect estimation using machine learning for healthcare application: tutorial and benchmark internet. arXiv 2021 [FREE Full text]
24. Glymour C, Zhang K, Spirtes P. Review of causal discovery methods based on graphical models. Front Genet 2019;10:524 [FREE Full text] [doi: 10.3389/fgene.2019.00524] [Medline: 31214249]
25. Vowels M, Camgoz N, Bowden R. D'ya like DAGs? A survey on structure learning and causal discovery. arXiv 2021 [FREE Full text] [doi: 10.1145/3527154]
26. Spirtes P, Glymour C. An algorithm for fast recovery of sparse causal graphs. Social Sci Comput Rev 2016 Aug 18;9(1):62-72. [doi: 10.1177/089443939100900106]
27. Colombo D, Maathuis M. Order-independent constraint-based causal structure learning. arXiv 2012 [FREE Full text]
28. Runge J, Nowack P, Kretschmer M, Flaxman S, Sejdinovic D. Detecting and quantifying causal associations in large nonlinear time series datasets. Sci Adv 2019 Nov;5(11):eaau4996 [FREE Full text] [doi: 10.1126/sciadv.aau4996] [Medline: 31807692]
29. Runge J. Recent progress and new methods for detecting causal relations in large nonlinear time series datasets. In: Proceedings of the EGU General Assembly 2020. 2020 Presented at: EGU General Assembly 2020; May 4–8, 2020; Online. [doi: 10.5194/egusphere-egu2020-9554]
30. Pearl J. Causality. Cambridge, England: Cambridge University Press; 2009.
31. Verma T. Graphical aspects of causal models. UCLA. 1992 Sep 1. URL: https://ftp.cs.ucla.edu/pub/stat_ser/r191.pdf [accessed 2021-02-15]
32. Spirtes P, Glymour C, Scheines R. Causation, Prediction, and Search. Cambridge, Massachusetts: The MIT Press; 2000.
33. Verma T, Pearl J. Equivalence and synthesis of causal models. In: Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence. 1990 Presented at: UAI '90: Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence; Jul 27 - 29, 1990; Cambridge, MA, USA URL: https://dl.acm.org/doi/10.5555/647233.719736
34. Escalante HJ, Escalera S, Guyon I, Baró X, Güçlütürk Y, Güçlü U, et al, editors. Explainable and Interpretable Models in Computer Vision and Machine Learning. Cham,Basel, Switzerland: Springer International Publishing; 2018.
35. Buntine W. Theory refinement on Bayesian networks. arXiv 2013 [FREE Full text] [doi: 10.1016/b978-1-55860-203-8.50010-3]

36. Andrews B, Ramsey J, Cooper GF. Scoring bayesian networks of mixed variables. Int J Data Sci Anal 2018 Aug 11;6(1):3-18 [FREE Full text] [doi: 10.1007/s41060-017-0085-7] [Medline: 30140730]
37. Chickering DM. Optimal structure identification with greedy search. J Mach Learn Res 2002;3:507-554 [FREE Full text]
38. Meek C. Causal inference and causal explanation with background knowledge. arXiv 2013:403-409 [FREE Full text] [doi: 10.1007/978-94-009-7731-0_8]
39. Ramsey J. Scaling up greedy causal search for continuous variables. arXiv 2015 [FREE Full text]
40. Cooper GF, Herskovits E. A Bayesian method for the induction of probabilistic networks from data. Mach Learn 1992 Oct;9(4):309-347. [doi: 10.1007/bf00994110]
41. Tsamardinos I, Brown LE, Aliferis CF. The max-min hill-climbing Bayesian network structure learning algorithm. Mach Learn 2006 Mar 28;65(1):31-78. [doi: 10.1007/s10994-006-6889-7]
42. Tsamardinos I, Aliferis C, Statnikov A. Time and sample efficient discovery of Markov blankets and direct causal relations. In: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. 2003 Presented at: KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining; Aug 24 - 27, 2003; Washington, D.C. [doi: 10.1145/956750.956838]
43. Shimizu S, Hoyer P, Hyvarinen A, Kerminen A. A linear non-Gaussian acyclic model for causal discovery. J Mach Learn Res 2006;7:2003-2030 [FREE Full text]
44. Shimizu S, Hyvarinen A, Kano Y, Hoyer P. Discovery of non-gaussian linear causal models using ICA. arXiv 2012 [FREE Full text]
45. Shimizu S, Inazumi T, Sogawa Y, Hyvarinen A, Kawahara Y, Washio T, et al. DirectLiNGAM: a direct method for learning a linear non-Gaussian structural equation model. arXiv 2011 [FREE Full text]
46. Hyvärinen A, Smith SM. Pairwise likelihood ratios for estimation of non-Gaussian structural equation models. J Mach Learn Res 2013 Jan;14(Jan):111-152 [FREE Full text] [Medline: 31695580]
47. Hoyer P, Janzing D, Mooij J, Peters J, Schölkopf B. Nonlinear causal discovery with additive noise models. In: Proceedings of the Advances in Neural Information Processing Systems 21 (NIPS 2008). 2008 Presented at: Advances in Neural Information Processing Systems 21 (NIPS 2008); Dec 8-11, 2008; Vancouver BC Canada.
48. Lee H, Danieletto M, Miotto R, Cherng S, Dudley J. Scaling structural learning with NO-BEARS to infer causal transcriptome networks. arXiv 2019.
49. Zheng X, Dan C, Aragam B, Ravikumar P, Xing E. Learning sparse nonparametric DAGs. arXiv 2019 [FREE Full text]
50. Yu Y, Chen J, Gao T, Yu M. DAG-GNN: DAG structure learning with graph neural networks. arXiv 2019 [FREE Full text]
51. Wei D, Gao T, Yu Y. DAGs with no fears: a closer look at continuous optimization for learning Bayesian networks. arXiv 2020 [FREE Full text]
52. Ng I, Zhu S, Chen Z, Fang Z. A graph autoencoder approach to causal structure learning. arXiv 2019 [FREE Full text]
53. Lachapelle S, Brouillard P, Deleu T, Lacoste-Julien S. Gradient-based neural DAG learning. arXiv 2019 [FREE Full text]
54. Goudet O, Kalainathan D, Caillou P, Guyon I, Lopez-Paz D, Sebag M. Learning functional causal models with generative neural networks. arXiv 2018.
55. Kalainathan D, Goudet O, Guyon I, Lopez-Paz D, Sebag M. Structural agnostic modeling: adversarial learning of causal graphs. arXiv 2022 [FREE Full text]
56. Zhu S, Ng I, Chen Z. Causal discovery with reinforcement learning. arXiv 2020 [FREE Full text]
57. Ramsey J, Glymour M, Sanchez-Romero R, Glymour C. A million variables and more: the Fast Greedy Equivalence Search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. Int J Data Sci Anal 2017 Mar;3(2):121-129 [FREE Full text] [doi: 10.1007/s41060-016-0032-z] [Medline: 28393106]
58. Zheng X, Aragam B, Ravikumar P, Xing E. DAGs with NO TEARS: continuous optimization for structure learning. arXiv 2018 [FREE Full text]
59. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al. TensorFlow: a system for large-scale machine learning. In: Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation is sponsored by USENIX. 2016 Presented at: 12th USENIX Symposium on Operating Systems Design and Implementation is sponsored by USENIX; Nov 2–4, 2016; Savannah, GA, USA.
60. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: an imperative style, high-performance deep learning library. arXiv 2019 [FREE Full text]
61. Niu W, Huang X, Xu K, Jiang T, Yu S. Pairwise interactions among brain regions organize large-scale functional connectivity during execution of various tasks. Neuroscience 2019 Aug 01;412:190-206. [doi: 10.1016/j.neuroscience.2019.05.011] [Medline: 31181368]
62. Butte AJ, Kohane IS. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. Pac Symp Biocomput 2000:418-429 [FREE Full text] [doi: 10.1142/9789814447331_0040] [Medline: 10902190]
63. Zhang K, Tian C, Zhang K, Johnson T, Jiang X. A fast PC algorithm with reversed-order pruning and a parallelization strategy. arXiv 2021 [FREE Full text]

64. Tu R, Zhang C, Ackermann P, Mohan K, Kjellström H, Zhang K. Causal discovery in the presence of missing data. arXiv 2020.
65. Huang B, Zhang K, Zhang J, Sanchez-Romero R, Glymour C, Schölkopf B. Behind distribution shift: mining driving forces of changes and causal arrows. Proc IEEE Int Conf Data Min 2017 Nov;2017:913-918 [FREE Full text] [doi: 10.1109/ICDM.2017.114] [Medline: 31068766]

Abbreviations

CPDAG: completed partially directed acyclic graph
DAG: directed acyclic graph
DAG-GNN: Directed Acyclic Graph-Graph Neural Network
FCI: Fast Causal Inference
FDR: false discovery rate
FPR: false positive rate
GAE: graph autoencoder
GES: Greedy Equivalence Search
GFCI: Greedy Fast Causal Interface
IC: inductive causation
LiNGAM: linear non-Gaussian acyclic model
MMHC: Max-Min Hill Climbing
PC: Peter Spirtes and Clark Glymour
SEM: structural equation model
SHD: structural hamming distance
TPR: true positive rate

Edited by C Lovis; submitted 25.03.22; peer-reviewed by M Banf, R Iyer; comments to author 12.06.22; revised version received 30.08.22; accepted 18.09.22; published 17.01.23.

Please cite as:
Upadhyaya P, Zhang K, Li C, Jiang X, Kim Y
Scalable Causal Structure Learning: Scoping Review of Traditional and Deep Learning Algorithms and New Opportunities in Biomedicine
JMIR Med Inform 2023;11:e38266
URL: https://medinform.jmir.org/2023/1/e38266
doi:10.2196/38266
PMID:36649070

Original Paper

Dealing With Missing, Imbalanced, and Sparse Features During the Development of a Prediction Model for Sudden Death Using Emergency Medicine Data: Machine Learning Approach

Xiaojie Chen[1*], MS; Han Chen[2*], PhD; Shan Nan[1], PhD; Xiangtian Kong[3], PhD; Huilong Duan[1,4], PhD; Haiyan Zhu[2,5], PhD

[1]Key Laboratory of Biomedical Engineering of Hainan Province, School of Biomedical Engineering, Hainan University, Haikou, China

[2]Hainan Hospital of Chinese People's Liberation Army General Hospital, Sanya, China

[3]IMWare, Wuhan, China

[4]College of Biomedical Engineering and Instrumental Science, Zhejiang University, Hangzhou, China

[5]First Medical Center of Chinese People's Liberation Army General Hospital, Beijing, China

[*]these authors contributed equally

Corresponding Author:
Haiyan Zhu, PhD
First Medical Center of Chinese People's Liberation Army General Hospital
28 Fuxing Road, Haidian District
Beijing, 100037
China
Phone: 86 13521361644
Email: xiaoyanzibj301@163.com

Abstract

Background: In emergency departments (EDs), early diagnosis and timely rescue, which are supported by prediction modes using ED data, can increase patients' chances of survival. Unfortunately, ED data usually contain missing, imbalanced, and sparse features, which makes it challenging to build early identification models for diseases.

Objective: This study aims to propose a systematic approach to deal with the problems of missing, imbalanced, and sparse features for developing sudden-death prediction models using emergency medicine (or ED) data.

Methods: We proposed a 3-step approach to deal with data quality issues: a random forest (RF) for missing values, k-means for imbalanced data, and principal component analysis (PCA) for sparse features. For continuous and discrete variables, the decision coefficient R^2 and the κ coefficient were used to evaluate performance, respectively. The area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPRC) were used to estimate the model's performance. To further evaluate the proposed approach, we carried out a case study using an ED data set obtained from the Hainan Hospital of Chinese PLA General Hospital. A logistic regression (LR) prediction model for patient condition worsening was built.

Results: A total of 1085 patients with rescue records and 17,959 patients without rescue records were selected and significantly imbalanced. We extracted 275, 402, and 891 variables from laboratory tests, medications, and diagnosis, respectively. After data preprocessing, the median R^2 of the RF continuous variable interpolation was 0.623 (IQR 0.647), and the median of the κ coefficient for discrete variable interpolation was 0.444 (IQR 0.285). The LR model constructed using the initial diagnostic data showed poor performance and variable separation, which was reflected in the abnormally high odds ratio (OR) values of the 2 variables of cardiac arrest and respiratory arrest (201568034532 and 1211118945, respectively) and an abnormal 95% CI. Using processed data, the recall of the model reached 0.746, the F_1-score was 0.73, and the AUROC was 0.708.

Conclusions: The proposed systematic approach is valid for building a prediction model for emergency patients.

(JMIR Med Inform 2023;11:e38590) doi:10.2196/38590

KEYWORDS

emergency medicine; prediction model; data preprocessing; imbalanced data; missing value interpolation; sparse features; clinical informatics; machine learning; medical informatics

Introduction

In the emergency department (ED), early identification of high-risk patients can improve clinical decisions, avoid waste of resources, and lead to better patient prognosis [1,2]. A prospective study showed that the incidence of adverse events due to improper emergency care is about 5%-10%, of which half can be prevented through early detection [3]. However, early identification is difficult as these patients often show little obvious signs before rapid deterioration [4].

Prediction models for high-risk patients in EDs can greatly support caregivers [5]. Electronic medical record (EMR) data, which fully capture patients' status, are an important source for developing disease risk prediction models [6]. As a typical high-risk disease in EDs, sudden death is a major public health problem worldwide, accounting for 15%-20% of all deaths [7,8]. A previous study showed that cardiogenic diseases, potassium, mean platelet volume, creatinine, chloride, and sodium are important variables to predict the risk of death in patients [5]. A survey showed that age, male, hypertension, diabetes, hypercholesterolemia, and a family history of coronary heart disease are all associated with increased risk of sudden death [9]. A study evaluating the relationship between the variables of laboratory tests and the occurrence of acute death in patients found that serum sodium, glucose, and the leukocyte count show a U-shaped relationship with mortality [10]. In addition, total bilirubin, creatine kinase, the international normalized ratio, aspartate aminotransferase, and lactate dehydrogenase are all risk factors associated with acute death in patients [11-13]. However, the data quality of EMRs limits their effective use for developing prediction models [6,14]. Prediction of sudden death needs a variety of clinical data, which are frequently missing, imbalanced, and having sparse features.

Missing values, imbalanced data, and sparse features are 3 common problems of EMR data. Missing values indicate not enough data collected due to improper use of the hospital information system or other reasons [14]. Imbalanced data refer to the imbalanced distribution of negative and positive samples. This leads to more features of negative samples in the learning model, which is not suitable for the prediction of arbitrary patients [15,16]. Sparse features are zero features that are much larger than nonzero features and increase computing memory and reduce generalization ability [17,18]. Especially in small samples, a large amount of noise in sparse features makes model training impossible to converge. Therefore, tackling these quality issues of EMR data is an essential step to improve the predictive performance of machine learning (ML) models.

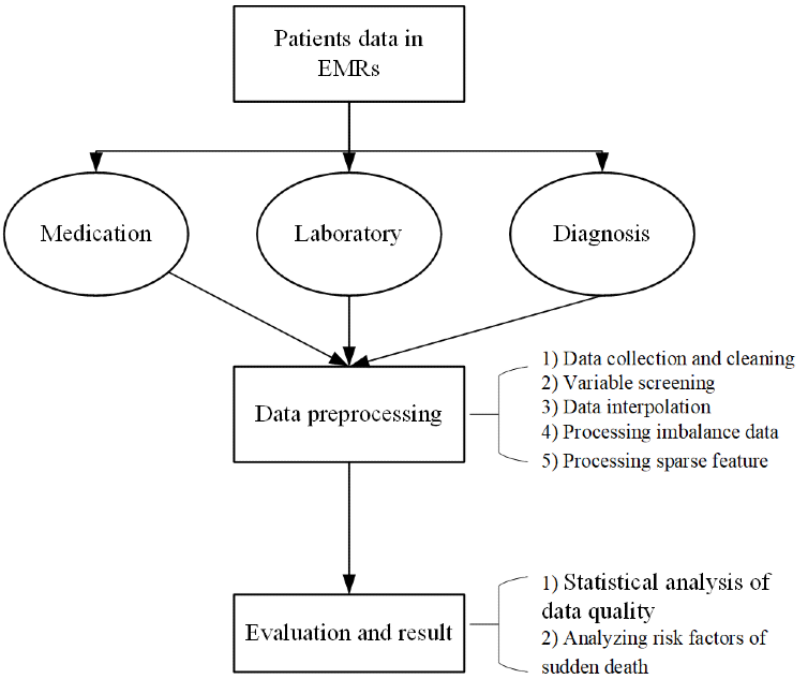
To solve the aforementioned 3 problems, we propose a series of ML approaches to increase fitting ability and generalization ability. Using the approach, we developed a sudden-death predication model. The risk factors related to sudden death obtained through logistic regression (LR) model were consistent with the results reported in the earlier literature on the analysis of risk factors of in-hospital death. These results show that our data-preprocessing approach can effectively maintain the rich information contained in emergency data and provide a reliable data source for the development of a sudden-death prediction model.

Methods

Study Design

Our methods of data preprocessing consisted of 5 steps, as shown in Figure 1. The last 3 steps tackle 3 low-quality issues: missing values, imbalanced data, and sparse features. Finally, postprocessing data quality is evaluated by a sudden-death prediction model case study.

Figure 1. Workflow of ED data preprocessing and evaluation. ED: emergency department; EMR: electronic medical record.



Data Collection and Cleaning

Data for ED patient prediction model development are summarized in Table 1.

Close investigation of each data table is required so as to know the location of our content of interest. For instance, data regarding a patient's basic information are stored in the emg_visit table. Lab test items and results are stored in the lab_result and lab_master tables. The clinical record field in the emg_order table can be used to determine whether a sudden-death event occurred. One lab test (eg, blood test) can be performed multiple times to observe the patient status closely. Based on clinical experts' opinions, only the last one is meaningful.

Table 1. Description of the data table involved in the query process.

Table name	Data description
emg_drug detail	The patient's medication record, including the prescription number, drug name, dosage, drug specification, administration time, and administration route during the treatment period
emg_drug_master	Master record form of patient medication recording patient ID and prescription number
emg_order	Doctor's order record form used to record the medication, inspection, diagnosis, treatment, and other doctor's orders of the patient during treatment
emg_visit	Patient visit information table, including the patient's basic personal information, diagnosis of the current visit, triage, and other information
lab_test_master	Patient's laboratory test master record form recording the patient's age and gender information, laboratory test items made during the visit, and the corresponding doctor's order ID
lab_result	Laboratory test results of patients, including test results of patients

Variable Screening

The number of variables obtained from the data collection was large, so screening of important variables facilitated final analysis. Two approaches can be adopted. One is based on statistical significance. The other is based on the specific research objective, opinions of medical experts, or authoritative literature [5,12,13]. In our study, the first approach was taken. Variables with many missing values were filtered out using the threshold. For example, Alvarez et al [19] set the threshold to 2%, while Seki et al [20] set it to 25%. In this study, we set the threshold to 80%. This means that when 80% of the values of 1 variable are missing, that variable should be filtered out.

Data Interpolation for Missing Values

Missing values affect the effectiveness of ML models. Data missing show 3 different patterns: missing completely at random (MCAR), missing at random deletion (MAR), and not missing at random (MNAR). MCAR means that the missing of data is completely random and does not depend on observed or unobserved values [21]. In this case, any interpolation method will not cause deviation. However, the assumption of MCAR in actual data is difficult to satisfy [22,23]. MNAR and MAR mean that the missing of data depends on the unobserved value and does not depend on the unobserved value, respectively [24]. However, it is impossible to infer whether the missing pattern belongs to MNAR or MAR through the existing data containing the missing pattern, and the assumption based on MAR is more consistent with the actual data situation [22,25]. MAR allows us to estimate missing values using existing observation data in the data set [24].

The goal of all kinds of interpolation methods is to reasonably estimate missing values and improve the quality of data. Interpolation methods are mainly divided into single interpolation and multiple interpolation. Multiple interpolation is a commonly used and better performance interpolation method. It generates multiple possible estimates for missing data and uses statistical inference to interpolate the final value. This method can reflect the randomness of missing data, and the interpolation error is smaller [21]. In a single interpolation, interpolation methods, such as constants (ie, specific identifications), mean, median, and data distribution, can be used. However, such methods usually cause greater deviation [26,27]. The single interpolation method based on ML has attracted increasingly more attention [23], such as interpolation based on a clustering algorithm [28], an ensemble model [29], and Bayesian theory [30]. Although multiple imputation can bring smaller deviation, when the frequency of missing data is high and the sample size is small, multiple imputation should be considered [31]. However, its implementation is relatively complex, and it needs to involve the selection of an interpolation model and the number of interpolation data created [32]. When the data are sufficient and the variability of the estimated value does not need to be considered, it is feasible to choose multiple imputation or single imputation [31]. Considering that our sample size was relatively sufficient, to build a simpler interpolation method, we used a random forest (RF) [33,34] as the interpolation algorithm to realize the interpolation of missing data in the form of a single interpolation.

Altogether, the followed steps are proposed.

- For variable "i," 1 set of patient samples without missing values work as training samples and the other set of patient samples with missing values work as test samples.
- If other variables in the 2 samples are missing, the mean (continuous variable) or mode (discrete variable) is temporarily interpolated to form a complete sample.
- Use training samples to train RF models, the model is applied to test samples to predict missing values.
- For the next variable, steps 1, 2, and 3 are repeated until all variables of the whole sample are interpolated.

Processing Imbalanced Data

Imbalanced data refer to the imbalanced distribution of negative and positive samples. For example, in the classification of rare diseases and credit predictions, there could be more negative samples than positive ones. Because most ML algorithms assume that categories (eg, positive or negative) of samples are evenly distributed, classifying models trained with imbalanced data are more likely to classify a new sample into the majority category [15].

Basic solutions for imbalanced data are to use under- or oversampling to make the data balanced, such as random oversampling [35], random undersampling, the synthetic minority oversampling technique (SMOTE) [36], and the adaptive synthetic sampling method (ADASYN) [15]. Although both undersampling and oversampling approaches can achieve data balance, the oversampling approach adds many sample copies to overfit the model. Wang and Japkowicz [16] and Chawla et al [36] also argued that undersampling is more favorable than oversampling in extreme imbalance situations. However, randomly discarding undersampling may also lose some representative samples. Segura-Bedmar et al [37] and Lin et al [38] proposed a clustering method to tackle this problem. The k-means considers the similarity between samples and uses the sample closest to the centroid of the cluster to approximate all the sample characteristics within the cluster, and the obtained samples are representative. The advantage of the clustering method over random undersampling is that all samples are used in the clustering process. This ensures that the information about all samples can be used to determine the sampling results and some important samples are not randomly discarded. In addition, we can adjust the number of clusters in k-means according to the actual data imbalance so as to achieve different undersampling ratios without other complex adjustments.

To avoid the loss of important samples, we adopted k-means based on the Euclidean distance to cluster samples. New samples were generated though clustering, which had similar characteristics in the same cluster and were distinguished in the different clusters. The centroid of a cluster represents the overall characteristics of the whole cluster. In this way, important features are not discarded. Since the centroid of the cluster is calculated based on the average of the samples in the cluster, the centroid is not necessarily a real sample. So, we took the real samples with the smallest distance from the centroid.

Processing Sparse Features

Sparse features means that the feature index is much larger than the actual number of nonzero features. In total, there were 891 different types of diagnosis in our data set. However, for a single patient, the number of diagnoses was quite few. This formed sparse-feature phenomena.

When sparse features occur, the sample is prone to having the problem of variable separation and multicollinearity. That is, a single variable or a linear combination of multiple variables can perfectly predict outcome events. However, this only works for small-size samples. It also leads to the situation in which the model gives an abnormally large weight to the variables and the results are unreliable [17,18,39]. Although there are many methods to optimize weights, such as gradient descent, a large number of zeros in features make the gradient tend to 0, and the parameters cannot be fully trained.

The processing of sparse features can be considered from both the model and the data themselves. From the point of view of the model, the parameter estimation bias of high-dimensional sparse data can be reduced through the optimization of the algorithm. For example, Firth regression [40] is used. The basic idea is to add a penalty term to the score function so as to reduce the deviation of the maximum-likelihood estimate of the parameter. This can solve the problem of variable separation and multicollinearity caused by sparse features to a large extent. From the point of view of the data themselves, it is necessary to transform the data to be processed into nonsparse data, and this transformation should retain the amount of information contained in the original data as much as possible. Considering the theme of our paper, our goal is to improve the quality of data rather than optimize the model algorithm. Therefore, we solved the problem of sparse features from the perspective of data. At present, there are many dimensionality reduction methods for high-dimensional sparse features, such as principal component analysis (PCA) [39], singular value decomposition (SVD) [41], and linear discriminant analysis (LDA) [42]. The essence of these methods is to map the original data to a low-dimensional space through a specific transformation form to solve the problem of data sparsity. Among these methods, LDA needs to reduce dimensionality based on sample labels. Considering that the actual data may not be able to carry labels, and the difference in label definitions will greatly affect the dimensionality reduction results, this supervised dimensionality reduction method is not conducive to being extended to other data scenarios [43]. Therefore, we considered using unsupervised dimensionality reduction methods, such as PCA, to transform our data.

PCA has been widely used in analysis with high-dimensional sparse features [44-46]. PCA essentially transforms the feature space of the original sample so that the new feature is a linear combination of the original features. The basic principle of principal component (PC) selection is to keep the maximum variance, and all PCs are orthogonal to one another. Thus, the phenomenon of multicollinearity is avoided. Therefore, new samples no longer have sparse features, which makes the ML model better fit the parameters.

In detail, new data can replace the original data as the input source for regression or classification models. Suppose  where each column represents a feature and each row is a sample. Assuming that the sample has been decentralized,  represents the covariance of matrix X. Let the transformed matrix Y = XV be D, which is derived as:

As C is a real symmetric matrix, according to the properties of the real symmetric matrix, its order m must have m unit orthogonal eigenvectors. That is,  is a matrix that can make the original covariance matrix similar to diagonalization.

Therefore, by solving m eigenvalues and eigenvectors of . By sorting the eigenvalues from large to small, we got $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_m)$. There are the following relationships:

Take the first k columns of V as the basis for transforming m-dimensional features into k-dimensional features and record it as the transformed sample matrix is Y = XP.

First, we manually merged similar diagnostic nouns according to prior knowledge, from 891 to 405. However, the data were obviously separated and sparse. For instance, none of the negative samples had a sudden cardiac arrest or sudden respiratory arrest diagnosis. Next, we only kept the diagnosis that appeared in more than 5% population. Finally, PCA was proposed for the remaining variables. The first 17 PCs that could explain 98.2% variance of the original sample were selected. Regression analysis was carried out on the samples after dimensionality reduction. The explanation of variables was achieved by counting the weight of the original variables on each PC.

Ethical Considerations

After preliminary review, the project was found to be in line with relevant medical ethics requirements. If it is funded by the Hainan Major Science and Technology Program in 2020, the Hainan Medical Ethics Committee will perform its duties and strictly abide by relevant regulations and requirements for medical ethics and informed consent of patients to ensure ethical supervision and review during the implementation of the project (reference number: 00824482406).

Results

Data Preprocessing and Model Building

A comprehensive evaluation was carried out on the ED data set of the Hainan Hospital of Chinese PLA General Hospital. We developed a set of Python programs to implement our methods. Specifically, the program was developed in Microsoft Windows 10 (Intel (R) core (TM) i5-9500 CPU, 3GHz). All data preprocessing and model building were completed in Python (Python 3.8 Anaconda) using multiple Python data science libraries, mainly including Numpy, Pandas, Matplotlib, and Scikit-learn. In addition, codes on data interpolation, imbalance correction, and PC regression are currently available on GitHub [47].

Data Collection and Cleaning

We collected the data of patients who went to the ED of the Hainan Hospital of Chinese PLA General Hospital from July 27, 2017, to May 6, 2021. In the sudden-death group, the data of 1085 patients were collected. In the non-sudden-death group, the data of 17,959 patients were collected. For the analysis of laboratory test data, we excluded patients who did not have any laboratory test records before sudden death. A total of 108 (10%) patients were excluded, and 977 (90%) patients with sudden death were used for the analysis of laboratory test data. For diagnostic data, we excluded patients who were missing diagnostic data from the visit. Finally, there were 1083 patients with sudden death and 615 patients with nonsudden death. We developed statistics on the baseline data of all patients, as shown in Supplementary Table S1 in Multimedia Appendix 1. Distributions of age and gender are visualized in Figures 2-5.

In the first group, there were 741 males (68.4%) and 342 females (31.6%), and 2 (0.2%) patients lacked gender information (Figure 2). The age varied between 45 and 80 years. The mean age was 56.4 years (SD 11.2). The quartile, median, and mode were 44, 59, and 68, respectively. In the second group, there were 9403 (52.4%) males and 8556 (47.6%) females. The age distribution is shown in Figures 4 and 5. The mean age was 41.6 years (SD 13.6). The quartile, median, and mode were 29, 42, and 48, respectively. For both groups, their distributions of age were akin to the normal distribution, which is consistent with a real-life situation.

Figure 2. Distribution of the gender of patients with sudden death.

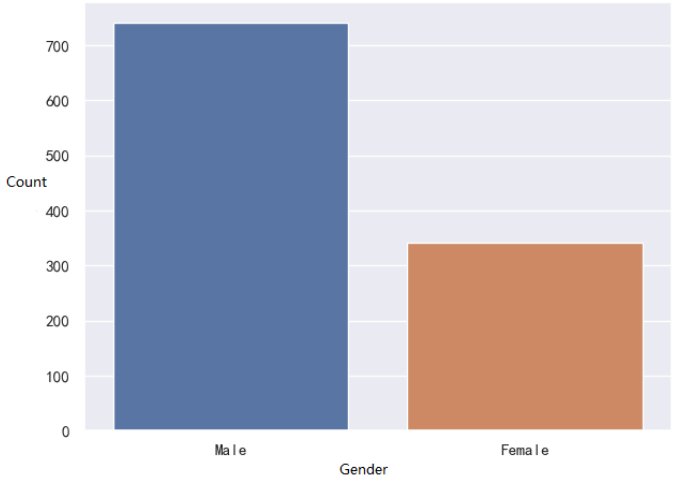


Figure 3. Distribution of the gender of patients without sudden death.

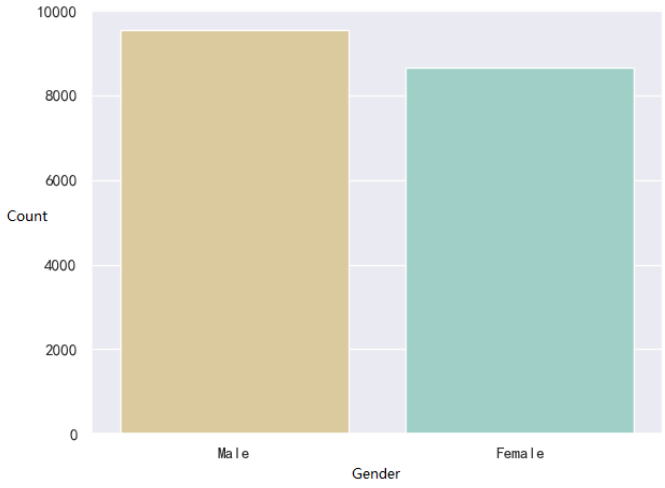


Figure 4. Distribution of age of patients with sudden death.

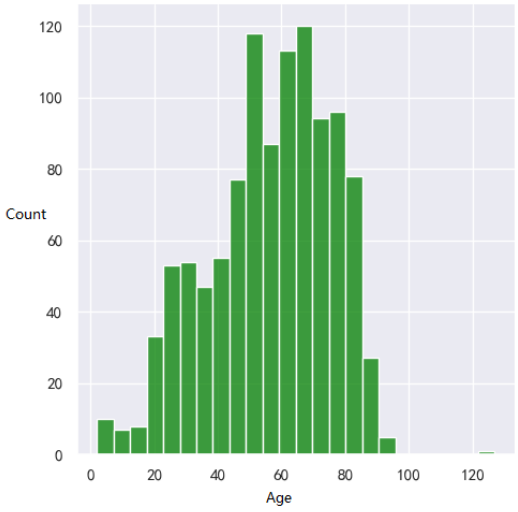
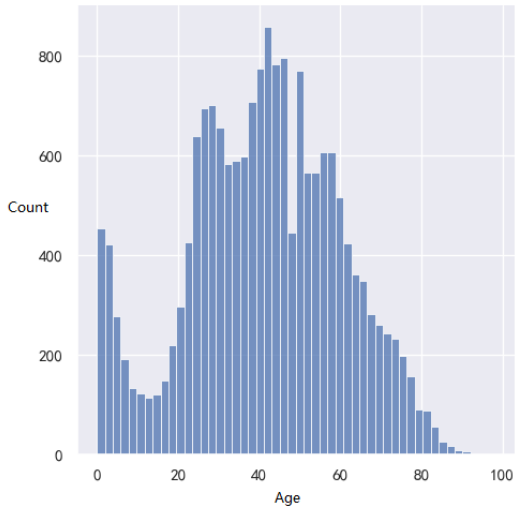


Figure 5. Distribution of patients of age with nonsudden death.



Variable Screening

To perform variable screening, that is, filtering out insignificant variables, we counted the total number of appearance and missing times. The second row of Table 2 shows the number of patients who had no corresponding data in the individual category. Moreover, we investigated the reasons missing data exist in all the 3 categories. For instance, there were 108 (10%) patients having no laboratory test. Among them, we could not find lab test data for 33 (30.6%) patients. For the remaining 75 (69.4%) patients, their lab tests appeared after the sudden-death event. There were 287 (26.4%) patients having no medication data. Sudden death had occurred before the medication was given, and the medication was in the doctor's order record, such as an epinephrine injection, but was not recorded in the patient's medication table.

There were 275 variables in the lab test category. For a given variable, not every patient (sample) had the value, namely a missing value. The missing ratio of a variable could be obtained by the number of cases having a missing value of that variable being divided by the total number of patients. The average ratio was 79.8%, as shown in the third row of Table 2. So, we set an 80% threshold to screen nonstatistically significant variables. Finally, 72 variables were kept in this category. These were patient age, gender, glucose, creatine kinase, inorganic phosphorus, total cholesterol, triglycerides, potassium, sodium, and calcium.

For diagnosis, 891 different types of diagnosis were obtained after the initial data collection. Because the diagnosis is recorded in the form of free text, 1 diagnosis item could have several different synonyms. By merging these texts into a unified name via manual review, we obtained 405 variables. The number of confirmed patients of each diagnostic variable was counted. Instead of an 80% threshold, 5% was considered. Considering both positive and negative samples, 18 diagnostic variables were kept. Among them, 11 (61.1%) variables were shared by both. These were myocardial infarction, chest distress, sudden cardiac arrest, fever, rib fracture, renal dysfunction, chest pain, diabetes, abdominal pain, pulmonary infection, respiratory arrest, trauma, atrial fibrillation, disturbance of consciousness, cerebral hemorrhage, cerebral infarction, coronary heart disease, and hypertension.

Table 2. Missing value ratios of variables of patients with sudden death.

	Laboratory tests (275 variables)	Medications (402 variables)	Diagnosis (891 variables)
Patients without data, n (%)	108 (10%)	287 (26.4%)	2 (0.18%)
Average ratio of missing values	79.8% (866/1085)	72.4% (786/1085)	99% (1080/1085)
Maximum ratio of missing values	90% (977/1085)	73.5% (797/1085)	100% (1085/1085)
Minimum ratio of missing values	25.8% (280/1085)	48.5% (526/1085)	58.4% (634/1085)

Data Interpolation, Processing Imbalanced Data, and Sparse Features

In addition to age and gender, we used an RF to interpolate the missing values for each of the remaining variables. Nonmissing patient data were used as a training set to train the model to interpolate missing values. The training set was further split into training data (80%) and validation data (20%). The coefficient of determination R^2 and the κ coefficient were used to test the consistency of the imputation results of continuous variables and categorical variables. In the interpolation process, the median of R^2 was 0.623 (IQR 0.647) and the median of the κ coefficient was 0.444 (IQR 0.285).

Due to the extreme imbalance of our original data, the number of patients with sudden death only accounted for 5% (977/18,936) of the total sample size. We generated 4 different data ratios (1:10, 1:5, 1:2, and 1:1) through k-means to achieve undersampling. These data were used with the original ratio to evaluate models of different data ratios and then to verify the rationality of our sampling method.

Validation by a Sudden-Death Case Study

Analyzing Risk Factors of Sudden Death

We constructed an LR model to analyze the patients' laboratory test variables using a data set with a data ratio of 1:1 as the data source to filter variables. To reflect the degree of correlation between variables, continuous variables were treated as ordinal categorical variables. Taking the normal index range of the variables as a reference point, the test results of the patients were mapped into 3 categories: L (index is lower than the normal value), N (index is normal), and H (index is higher than the normal value). To determine the significant factors affecting the sudden death of patients and avoid a negative effect on the final analysis results, we first performed the chi-square test to filter out the variables and then excluded variables when $P>.10$. Next, LR univariate analysis was performed to filter out variables with $P>.05$. Tables 3 and 4, respectively, show the variables excluded by the chi-square test and the LR univariate analysis, and their P values. We reintroduced some of the excluded variables into the final candidate variable set according to the literature review and the advice of consulting medical experts, including urine specific gravity, chloride, hematocrit, sodium, magnesium, lactate dehydrogenase, urine ketone body test, red blood cell count, and serum albumin. These variables have no significant statistical significance but are clinically related to sudden death. Finally, we selected 4 subgroups from the set of variables with significant statistical significance. In addition, variables not statistically significant but related to outcome events were also grouped separately. The final 5 groups were subjected to LR multivariate analysis, and the groups were as follows:

- Group 1: qualitative test of creatinine, serum uric acid, urine protein
- Group 2: γ-glutamyl transferase, alanine aminotransferase, total bilirubin

- Group 3: international normalized ratio, platelet count, plasma prothrombin time
- Group 4: potassium, creatine kinase
- Group 5: urine specific gravity, chloride, hematocrit, sodium, magnesium, lactate dehydrogenase, urine ketone body test, red blood cell count, serum albumin

For each group, 500-fold bootstrapping was used for model training and evaluation [48]. Each bootstrap randomly split 70% of the data into the training set and 30% of the data into the test set. Finally, the mean values of AUROC, recall, and F_1-score for 500 training sessions in each group were reported, and the AUROC also reported the 95% CI. Table 5 illustrates the model evaluation results of the 5 groups of variables. The performance parameters of group 2 were the best among the 5 groups of variables. In the recognition of patients with sudden death, a recall rate of 0.801 was obtained, the F_1-score was 0.835, and the model's AUROC was 0.843 (95% CI 0.842-0.844). The results showed that this set of variables can better identify patients with sudden death. Therefore, other group variables based on the group 2 variables were added successively, and AUROC was taken as the evaluation index. The added variables would be included in the final model if AUROC could be improved. In the end, 13 laboratory test risk variables related to sudden death events were determined, and the patient's gender variable was retained as a demographic feature. In general, the final variables used included γ-glutamyl transferase, alanine aminotransferase, total bilirubin, creatinine, serum uric acid, the international standardized ratio, creatine kinase, the platelet count, potassium, sex, sodium, magnesium, chloride, and serum albumin. These variables were used to build the final LR model. Table 6 shows the results of LR multivariate analysis.

After determining the patient features for analysis, we split the original scale data into a training set (70%) and a test set (30%). For the training set, 4 different categories of data sets (1:1, 1:2, 1:5, 1:10) were formed by undersampling to train the model. Finally, the performance of the model was evaluated on the test set. The mean and 95% CI (500-fold bootstrapping) of the final AUROC, AUPRC, F_1-score, and recall are shown in Supplementary Table S2 in Multimedia Appendix 1. In addition, we further used Brier scores to evaluate the calibration ability of models trained with different data ratios.

In general, as the data ratio tended to balance, the performance of the model gradually improved. Figures 6 and 7 show the model receiver operating characteristic (ROC) curve (Figure 6) and the precision-recall (PR) curve (Figure 7) of the 4 data ratios. In recognizing patients with sudden death, the best model obtained a recall rate of 0.863 (95% CI 0.862-0.865), the F_1-score was 0.84 (95% CI 0.839-0.842), the AUROC of the model was 0.895 (95% CI 0.894-0.896), and the AUPRC was 0.897 (95% CI 0.896-0.899). The original scale data model performed the worst, with an AUROC of 0.812 (95% CI 0.811-0.813) and an AUPRC of 0.407 (95% CI 0.404-0.409). We plotted the reliability curves of 5 training sets with different data ratios on the same test set and calculated Brier scores (Supplementary Figure S1 in Multimedia Appendix 1). Consistent with the viewpoint mentioned by Geeven et al [49], imbalance correction actually weakened the clinical application value of the model, which was mainly manifested in the poor calibration ability of the model. With the increase in sampling, the calibration of the model was worse and the Brier score was 0.16 and 0.108 in the data ratio of 1:1 and the original data ratio, respectively. Imbalance correction can balance the sensitivity and specificity of the model to a greater extent and avoid biased errors in the model. Undersampling optimizes the AUROC, F_1-score, and AUPRC of the model trained by the proportion of the original data. Although the Brier score in calibration improved, the gap was not large. To observe the risk factors of sudden death in patients more intuitively, we visualized the regression coefficients of the best model after performing LR(Figure 8) to observe the relationship between variables and sudden-death events.

Table 3. Statistics of variables filtered by the chi-square test.

Variable	χ^2 (*df*)	*P* value
Monocytes	5.433 (6)	.49
Basophil	0.705 (4)	.95
Eosinophils	0.977 (4)	.91
Urine specific gravity determination	0 (2)	.99
Urine tube type	1.25 (4)	.87
Urine tube type (microscopic examination)	6.863 (8)	.98
Qualitative test of urinary bilirubin	13.185 (4)	.21
Mean erythrocyte hemoglobin concentration	7.828 (6)	.25
Chloride	4.649 (6)	.59
Erythrocyte volume distribution width measurement coefficient of variation (CV)	1.148 (4)	.89
Hematocrit assay	4.982 (6)	.55
Sodium	7.915 (6)	.24
Magnesium	10.22 (6)	.12

Table 4. Statistics of variables screened by LR[a] univariate analysis.

Variable	Reference range	OR[b] (95% CI)	*P* value
Lactate dehydrogenase	50.0-150.0 U/L	1.029 (0.94-1.127)	.53
Urine ketone body test	N/A[c]	0.912 (0.769-1.081)	.29
Red blood cell count	3.5-5.9 1012/L	0.827 (0.642-1.065)	.14
Serum albumin	35.0-50.0 g/L	0.893 (0.689-1.157)	.39
High-density lipoprotein cholesterol	1.0-1.6 mmol/L	0.961 (0.749-1.232)	.75

[a]LR: logistic regression.

[b]OR: odds ratio.

[c]N/A: not applicable.

Table 5. Comparing the performance of 5 groups of variables.

Group	Recall	F_1-score	AUROC[a] (95% CI)
1	0.478	0.6	0.683 (0.681-0.684)
2	0.801	0.835	0.843 (0.842-0.844)
3	0.606	0.687	0.725 (0.724-0.727)
4	0.484	0.605	0.686 (0.685-0.687)
5	0.852	0.651	0.562 (0.561-0.564)

[a]AUROC: area under the receiver operating characteristic curve.

Table 6. LR[a] multivariate analysis.

Variable	Reference range	OR[b] (95% CI)
γ-Glutamyl transferase	0.0-50.0 U/L	0.225 (0.222-0.228)
Alanine aminotransferase	5.0-40.0 U/L	1.828 (1.804-1.852)
Total bilirubin	0.0-21.0 μmol/L	19.954 (19.7-20.2)
Creatinine	30.0-110.0 μmol/L	1.352 (1.331-1.372)
Serum uric acid	104.0-444.0 μmol/L	1.346 (1.334-1.359)
International normalized ratio	0.8-1.2	2.23 (2.188-2.272)
Creatine kinase	24.0-320.0 U/L	2.457 (2.431-2.483)
Platelet count	100.0-300.0 $\times 10^9$/L	0.623 (0.617-0.629)
Potassium	3.5-5.1 mmol/L	1.057 (1.043-1.07)
Gender	Female	0.183 (0.182-0.184)
Sodium	135-145 mmol/L	2.182 (2.102-2.262)
Magnesium	0.8-1.0 mmol/L	4.807 (4.587-5.027)
Chloride	96.00-106.00 mmol/L	0.615 (0.603-0.627)
Serum albumin	35-51g/L	1.284 (1.268-1.3)

[a]LR: logistic regression.

[b]OR: odds ratio.

Figure 6. ROC curves of different data ratio. AUC: area under the curve; ROC: receiver operating characteristic.

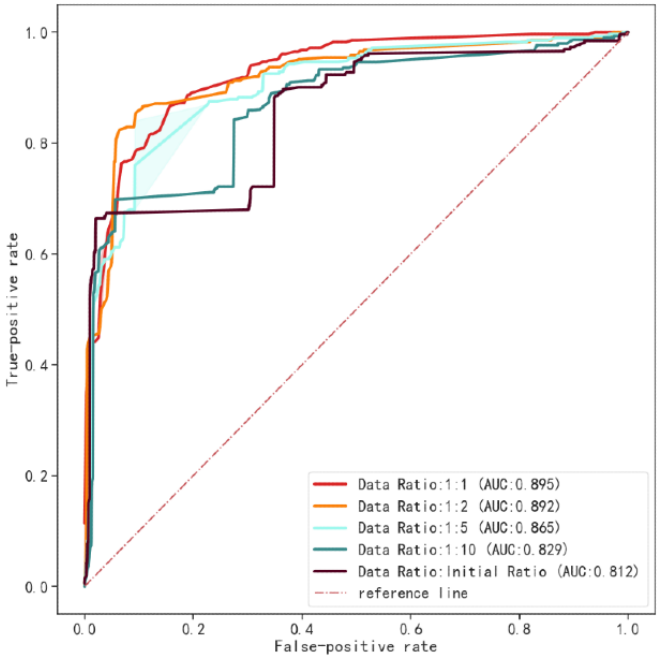


Figure 7. PR curves of different data ratio. AUPRC: area under the precision-recall curve; PR: precision-recall.

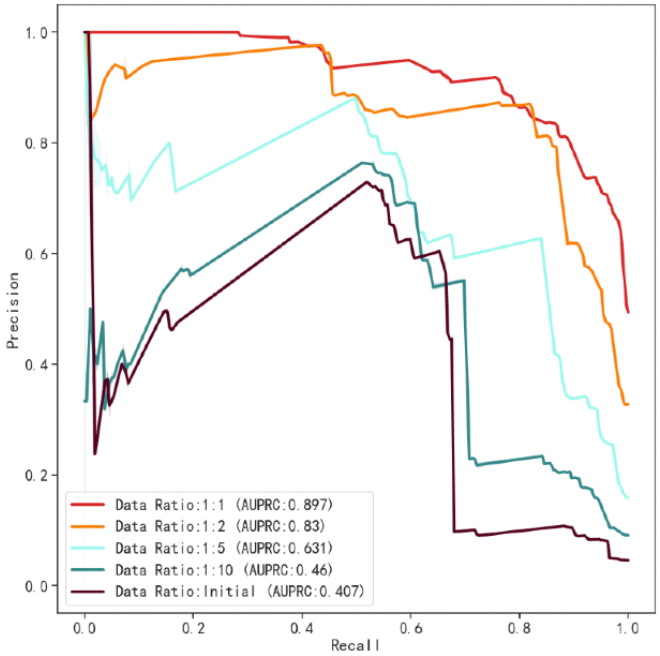
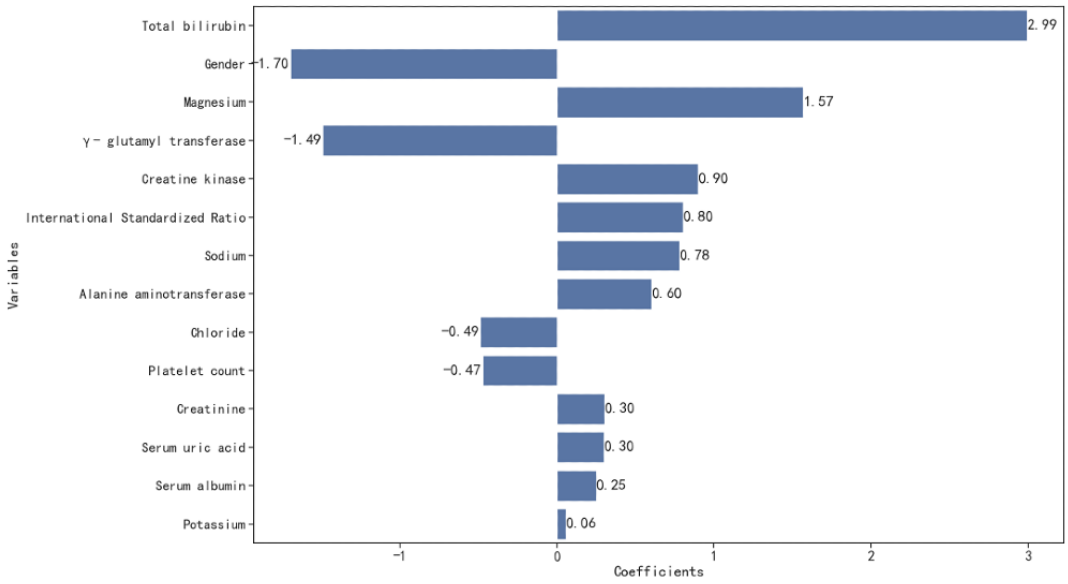


Figure 8. Visualization of logistic regression coefficients.



Development of Other ML Models

We use interpolated and undersampled data (data ratio 1:1) to train several other ML models and evaluate their performance. The training models included an RF [50], a gradient boosting machine (GBM) [51], a support vector machine (SVM) [52], and least absolute shrinkage and selection operator (LASSO) [53], which are also often used to develop medical prediction models [49,54]. We use 500-fold bootstrapping for internal validation. Each bootstrap used 70% data for training and the remaining 30% data for performance evaluation. The area under the curve (AUC), AUPRC, recall, and F_1-score and their 95% CI values were reported. Before model training, a grid search was conducted to tune the best hyperparameter of each model through 5-fold cross-validation. The hyperparameter settings of each model are shown in Supplementary Table S7 in Multimedia Appendix 1. The ROC curve and PR curve of the models are shown in Supplementary Figures S2 and S3 in Multimedia Appendix 1, respectively, and the performance evaluation results are shown in Supplementary Table S8 in Multimedia Appendix 1. In general, the performance of the RF and GBM with an integrated scheme was the best, with an AUC of 0.936 (95% CI 0.934-0.937) and 0.931 (95% CI 0.93-0.932),respectively, and an F_1-score of 0.857 (95% CI 0.856-0.858) and 0.821 (95% CI 0.82-0.823), respectively. This can benefit from the generalization and the ability to deal with complex feature relationships of the integrated model. The comprehensive decision results of multiple base learners are more stable than the single-model prediction results, and the performance is better. The SVM also performed better than the LR and LASSO, which are linear models, with an AUC of 0.913 (95% CI 0.912-0.914). This shows that there are some nonlinear features we used that made the linear model insufficient to recognize the relationship between these features.

Diagnostic Data Analysis Results

The final sample included 1083 patients with sudden death and 615 patients with nonsudden death. Table 7 shows the number of confirmed patients with 18 variables. The final diagnostic variables used included hypertension, myocardial infarction, cerebral hemorrhage, cardiac arrest, absolute pain, atmospheric fabric, fever, trauma, respiratory arrest, diabetes, corporate heart disease, and cerebral infarction.

We used 500-fold bootstrapping for internal validation of the model. For each bootstrap, 70% of the samples were randomly selected as the training set and 30% as the test set to evaluate the model. The final reported model performance was the mean and 95% CI of 500 results [48].

The first 17 PCs that could explain 98.2% of the variance of the original sample were selected as new variables for analysis. To observe the role of PCA, we compared the 2 schemes: the LR model using the original data and the LR model after dimensionality reduction using PCA. The LR model trained with the original data obtained a recall rate of 0.445 (95% CI 0.443-0.448), an F_1-score of 0.562 95% CI 0.56-0.564), and an AUROC of 0.602 (95% CI 0.6-0.603). After PCA dimensionality reduction of the original data, the PC variable was used as the data source to train the LR model, and a recall rate of 0.746 (95% CI 0.731-0.76) was obtained, the F_1-score was 0.73 (95% CI 0.721-0.738), and the AUROC of the model was 0.708 (95% CI 0.707-0.71). Figure 9 shows the ROC curves of the 2 models. The LR model using the original data had the phenomenon of variable separation, which is reflected in the abnormally high OR values of cardiac arrest and respiratory arrest (201568034532 and 1211118945) and an abnormal 95% CI, which makes the results unreliable. In addition, the performance of the model was poor, and only a recall rate of 0.445 was obtained in the identification of patients with sudden death, which means that the identification ability of the model for patients with sudden death is not strong. After PCA dimensionality reduction, the data were no longer sparse, the model parameters were better fitted, and the model performance improved to a certain extent. In addition, data conversion also eliminated the problems of variable separation and multicollinearity.

To determine the impact of various diagnostic variables on the sudden death of emergency patients, we statistically analyzed the results of multivariate analysis on 17 PCs input into the LR model. The OR of PC4, PC5, and PC6 was 3.044, 2.859, and 3.931, respectively, showing a significant correlation with sudden-death events (Table 8). In each PC, the magnitude of the loading, the elements in the PC, reflected the importance of the original variable in the PC (Supplementary Table S3 in Multimedia Appendix 1). The loadings of all components showed that cerebral infarction, hypertension, and pulmonary infection were the top 3 variables in PC4. In PC5 and PC6, the top 3 variables were consciousness disorder, diabetes, and fever. Based on the results of the 3 PCs, we believe that the 6 diagnoses of cerebral infarction, hypertension, pulmonary infection, consciousness disorder, diabetes, and fever are significantly associated with sudden death in emergency patients.

Table 7. Statistics of people diagnosed.

Variable	People with sudden death diagnosed, n (%)/people with nonsudden death diagnosed, n (%)
Myocardial infarction	57 (5.26)/23 (3.74)
Chest tightness	8 (0.74)/35 (5.69)
Cardiac arrest	120 (11.08)/0
Fever	50 (4.62)/43 (6.99)
Rib fracture	58 (5.36)/3 (0.49)
Abnormal renal function	42 (3.88)/35 (5.69)
Chest pain	18 (1.66)/38 (6.18)
Diabetes	65 (6.00)/66 (10.73)
Abdominal pain	30 (2.77)/45 (7.32)
Pulmonary infection	85 (7.85)/64 (10.41)
Respiratory arrest	106 (9.79)/0
Trauma	58 (5.36)/16 (2.60)
Atrial fibrillation	39 (3.60)/33 (5.37)
Disturbance of consciousness	82 (7.57)/17 (2.76)
Cerebral hemorrhage	77 (7.11)/26 (4.23)
Cerebral infarction	75 (6.93)/71 (11.54)
Coronary heart disease	29 (2.68)/39 (6.34)
Hypertension	65 (6.00)/106 (17.24)

Figure 9. ROC curves of 2 models. AUC: area under the curve; LR: logistic regression; PCA: principal component analysis; ROC: receiver operating characteristic.

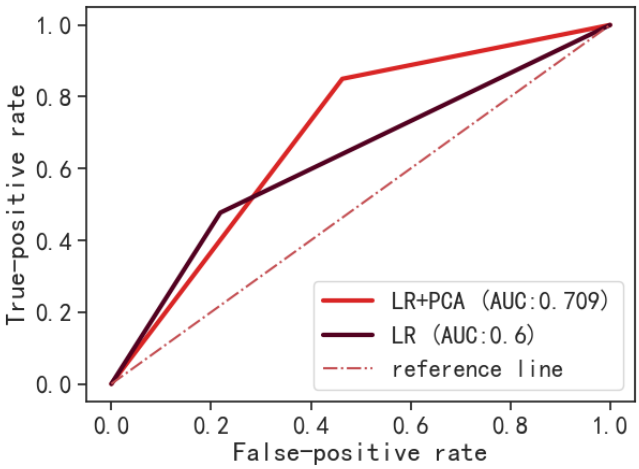


Table 8. PC[a] regression results

PC	OR[b] (95% CI)
1	0.239 (0.235-0.242
2	2.429 (2.383-2.476)
3	1.19 (1.126-1.253)
4	3.044 (2.948-3.141)
5	2.859 (2.687-3.031)
6	3.931 (3.714-4.148)
7	1.49 (1.405-1.575)
8	1.699 (1.562-1.836)
9	2.104 (1.949-2.259)
10	2.153 (2.016-2.289)
11	2.451 (2.191-2.711)
12	2.031 (1.855-2.206)
13	1.457 (1.339-1.575)
14	0.949 (0.863-1.034)
15	1.423 (1.231-1.614)
16	2.546 (2.221-2.871)
17	0.182 (0.164-0.201)

[a]PC: principal component.

[b]OR: odds ratio.

Discussion

Principal Findings

In this paper, 3 ML schemes were proposed to deal with missing, imbalanced, and sparse features in the process of developing sudden-death prediction models using emergency medicine data, which improves the performance of the developed model. To solve the problem of missing data, we propose an RF method to use real data to interpolate missing data. In the interpolation process, the consistency of the interpolation results is checked by determining the coefficient R^2 and the κ coefficient. From the interpolation results, the method shows the ability to correctly interpolate missing data. Imbalanced data are not conducive to obtaining accurate analysis results, and the model will be more inclined to predict new samples as patients with nonsudden death [15]. In view of this phenomenon, we used the k-means algorithm to generate multiple data sets with different proportions of different categories by undersampling to evaluate the model. The method based on k-means can better preserve the patient's characteristic information. This method will not lose some representative patient samples due to random discarding, thus reducing the bias caused by sampling. The results show that the comprehensive performance of the model gradually improves as the data tend to balance (Figures 3-5). However, imbalance correction will weaken the calibration ability of the model and increase the calibration error. Data sparsity is also not conducive to modeling and analysis. When the samples are too sparse, the results of the classifier based on maximum-likelihood estimation will become unreliable, because there may be variable separation and multicollinearity [18,55]. PC regression analysis is a method that uses PCA to extract the PC information about the original samples and uses PCs to replace the original variables for regression modeling [39]. In our diagnostic data, the LR model using the original data showed the phenomenon of variable separation, which led to unreliable results and poor performance. The performance of the PC regression model has been improved. In addition, we can analyze the diagnosis significantly related to the sudden death of emergency patients from the results of PC regression. These diagnoses are consistent with previous findings [9].

At present, there are many studies on the prediction of sudden death. Yu et al [54] constructed an ML model to predict sudden cardiac death (SCD) in 15,661 patients with atherosclerosis. The results showed that the ML model performs better than the standard Poisson regression model and the AUROC of the ML model was 0.89. Karen et al [56] trained an ML-based early warning model for identifying sudden infant death syndrome using the public data set “Lipidomic in sudden infant death syndrome.” The RF algorithm achieved an AUROC of 0.9 and a recall of 0.8. Ye et al [5] selected a variety of ML algorithms to build an early real-time early warning system (EWS) to predict the death risk of emergency patients and carried out prospective validation. The results showed that the EWS could give an early warning within 40 hours before sudden death, and the AUROC reached 0.884. Bhattacharya et al [57] used the electronic health records of 711 patients with hypertrophic myocardial cake and established an LR and naive Bayesian model with 22 variables, including statins, a family history of

SCD, and left ventricular ejection fraction, to predict the risk of sudden death (ventricular fibrillation) in these patients. The sensitivity and specificity of the optimal model were 0.73 and 0.76, respectively, and the AUROC was 0.83. For our model, in the LR model constructed by using laboratory test data, the AUROC reached 0.895. After imbalance correction, the recall rate and AUPRC improved, reaching 0.863 and 0.897, respectively. Compared to the existing sudden-death prediction model based on ML, the performance of our model can achieve a similar effect, further indicating that our data-preprocessing methods can preserve the patient's characteristic information and improve the availability of emergency care.

Limitations

This work also has some limitations. On the one hand, we only considered a single ML algorithm for data interpolation and did not discuss and compare the application of other possible ML algorithms in interpolation. It is possible that we overlooked the better performance of other methods. For example, for our data, due to the large proportion of missing and seriously imbalanced categorical variables, although we tried to adjust the relatively balanced data set to train the model, the κ coefficient improved to a certain extent but the effect was still poor. Therefore, a further discussion of ML methods that can handle a large number of missing and unbalanced categories or more reasonable feature processing may achieve better imputation results. Although imbalance correction can improve the sensitivity and specificity of the model, it can avoid biased errors of the model. However, this correction will also weaken the clinical application value of the model, lowering the calibration ability of the model and making it unable to accurately estimate the risk probability of patients. For the prediction model, the calibration ability of the model was not high, even on the original scale data set. Model calibration is another important characteristic of evaluating the clinical significance of prediction models. A well-calibrated model can provide more useful information for clinical decisions [58,59]. We can further consider using isotonic regression [60] to calibrate the model to improve its clinical application value. In addition, although the solution to deal with missing, imbalanced, and sparse features proposed by us is not the latest method, it is sufficient to solve the main data quality problems encountered in the development of prediction models for sudden death, which is reflected in the improvement of model performance and the consistency of the risk factors of sudden death obtained with the earlier literature results. In the future, we need to further explore the latest methods to solve these 3 data quality problems so as to extend the data-processing process to other data sets and provide a more reliable data source for prediction models. With regard to the construction of risk factor prediction models for patients with sudden death, we have a broad definition of sudden death, including patients who have undergone rescue or death events. These patients may include some nonemergency death cases, which may have a confusing effect on the final model. In addition, our feature selection was completely based on data, and only the remaining variables were trained in groups during the model training stage. This form can reduce the complexity of manually selecting features and also explore some potential risk variables. However, some clinically significant variables will also be discarded. Therefore, whether the model has clinical guiding significance remains to be further investigated. As a case study, we used LR as the main prediction model, which facilitated us to develop and analyze the risk factors of sudden death. However, the processing capacity of the LR model for nonlinear predictors is insufficient, resulting in insufficient performance of the developed model [17]. This can be seen from the results of other ML models we additionally developed (the RF and GBM had the best performance, with an AUC of 0.936 and 0.931, respectively, which are better than LR models). Therefore, in the future, we will further optimize the data-preprocessing process and try to develop ML models with better performance to improve the clinical usability.

Conclusion

Our work proposes to use ML methods to deal with data quality issues, such as missing data, data imbalance, and sparse features in emergency data, so as to improve data availability. In addition, the risk factors of sudden death in emergency patients are obtained from our model analysis. As a preliminary analysis result, this result is also the basis for the later use of ML algorithms to build the feature selection and data analysis of the prediction model of sudden death in emergency patients.

Acknowledgments

The authors would like to show their appreciation to the engineers working in the Information Centre (ICT) department of the Hainan Hospital of Chinese PLA General Hospital for their help with data preparation.

The publication of this paper was funded by grants from the National Natural Science Foundation of China (no. 82102187) and the Hainan Natural Science Foundation Youth Fund (no. 620QN380).

Data Availability

The data sets used and analyzed during this study are available from the first author upon reasonable request.

Authors' Contributions

XC carried out the methodological study and drafted the manuscript. HC collected and processed the data and drafted the manuscript. SN made the conceptual design and made critical revisions to the manuscript. XK reviewed the methodology and reviewed the manuscript. HD also reviewed the manuscript. HD conceptualized the study and performed a critical review.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary Tables and Figures.
[DOCX File , 278 KB - medinform_v11i1e38590_app1.docx]

References

1. Nagamine T, Gillette B, Pakhomov A, Kahoun J, Mayer H, Burghaus R, et al. Multiscale classification of heart failure phenotypes by unsupervised clustering of unstructured electronic medical record data. Sci Rep 2020 Dec 07;10(1):21340 [FREE Full text] [doi: 10.1038/s41598-020-77286-6] [Medline: 33288774]
2. Fernandes M, Mendes R, Vieira SM, Leite F, Palos C, Johnson A, et al. Risk of mortality and cardiopulmonary arrest in critical patients presenting to the emergency department using machine learning and natural language processing. PLoS One 2020;15(4):203 [FREE Full text] [doi: 10.1371/journal.pone.0230876] [Medline: 32240233]
3. Goulet H, Guerand V, Bloom B, Martel P, Aegerter P, Casalino E, et al. Unexpected death within 72 hours of emergency department visit: were those deaths preventable? Crit Care 2015 Apr 08;19(1):154-164 [FREE Full text] [doi: 10.1186/s13054-015-0877-x] [Medline: 25887707]
4. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLoS One 2017;12(4):174-186 [FREE Full text] [doi: 10.1371/journal.pone.0174944] [Medline: 28376093]
5. Ye C, Wang O, Liu M, Zheng L, Xia M, Hao S, et al. A Real-Time Early Warning System for Monitoring Inpatient Mortality Risk: Prospective Study Using Electronic Medical Record Data. J Med Internet Res 2019 Jul 05;21(7):137 [FREE Full text] [doi: 10.2196/13719] [Medline: 31278734]
6. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JPA. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J Am Med Inform Assoc 2017 Jan 17;24(1):198-208 [FREE Full text] [doi: 10.1093/jamia/ocw042] [Medline: 27189013]
7. Srinivasan NT, Schilling RJ. Sudden cardiac death and arrhythmias. Arrhythm Electrophysiol Rev 2018 Jun;7(2):111-117. [doi: 10.15420/aer.2018:15:2]
8. Lewis ME, Lin F, Nanavati P, Mehta N, Mounsey L, Nwosu A, et al. Estimated incidence and risk factors of sudden unexpected death. Open Heart 2016 Mar 23;3(1):32 [FREE Full text] [doi: 10.1136/openhrt-2015-000321] [Medline: 27042316]
9. Adabag AS, Luepker RV, Roger VL, Gersh BJ. Sudden cardiac death: epidemiology and risk factors. Nat Rev Cardiol 2010 Apr 9;7(4):216-225 [FREE Full text] [doi: 10.1038/nrcardio.2010.3] [Medline: 20142817]
10. Asadollahi K, Hastings I, Beeching N, Gill G. Laboratory risk factors for hospital mortality in acutely admitted patients. QJM 2007 Aug 02;100(8):501-507. [doi: 10.1093/qjmed/hcm055] [Medline: 17609227]
11. Du Z, Wei YG, Chen KF, Li B. An accurate predictor of liver failure and death after hepatectomy: a single institution's experience with 478 consecutive cases. World J Gastroenterol 2014 Jan 07;20(1):274-281 [FREE Full text] [doi: 10.3748/wjg.v20.i1.274] [Medline: 24415882]
12. Hernesniemi JA, Mahdiani S, Tynkkynen JA, Lyytikäinen LP, Mishra PP, Lehtimäki T, et al. Extensive phenotype data and machine learning in prediction of mortality in acute coronary syndrome - the MADDEC study. Ann Med 2019 Mar;51(2):156-163 [FREE Full text] [doi: 10.1080/07853890.2019.1596302] [Medline: 31030570]
13. Farahani NZ, Arunachalam SP, Sundaram DSB, Pasupathy K, Enayati M, Arruda-Olson AM. Explanatory analysis of a machine learning model to identify hypertrophic cardiomyopathy patients from EHR using diagnostic codes. Proc IEEE Int Conf Bioinformatics Biomed 2020 Dec;2020:1932-1937 [FREE Full text] [doi: 10.1109/bibm49941.2020.9313231] [Medline: 34316386]
14. Wells BJ, Chagin KM, Nowacki AS, Kattan MW. Strategies for handling missing data in electronic health record derived data. EGEMS (Wash DC) 2013 Dec;1(3):1035-1043 [FREE Full text] [doi: 10.13063/2327-9214.1035] [Medline: 25848578]
15. Chang HK, Wu CT, Liu JH, Lim WS, Wang HC, Chiu SL, et al. Early detecting in-hospital cardiac arrest based on machine learning on imbalanced data. IEEE Int Conf Healthc Informatics 2019:1-10. [doi: 10.1109/ICHI.2019.8904504]
16. Wang BX, Japkowicz N. Boosting support vector machines for imbalanced data sets. Knowl Inf Syst 2009 Mar 5;25(1):1-20. [doi: 10.1007/s10115-009-0198-y]
17. Heinze G. A comparative investigation of methods for logistic regression with separated or nearly separated data. Stat Med 2006 Dec 30;25(24):4216-4226. [doi: 10.1002/sim.2687] [Medline: 16955543]
18. Lee H, Park YM, Lee S. Principal component regression by principal component selection. Commun Stat Appl Methods 2015 Mar 31;22(2):173-180. [doi: 10.5351/csam.2015.22.2.173]
19. Alvarez CA, Clark CA, Zhang S, Halm EA, Shannon JJ, Girod CE, et al. Predicting out of intensive care unit cardiopulmonary arrest or death using electronic medical record data. BMC Med Inform Decis Mak 2013 Feb 27;13(1):28-35 [FREE Full text] [doi: 10.1186/1472-6947-13-28] [Medline: 23442316]

20. Seki T, Kawazoe Y, Ohe K. Machine learning-based prediction of in-hospital mortality using admission laboratory data: a retrospective, single-site study using electronic health record data. PLoS One 2021;16(2):246-255 [FREE Full text] [doi: 10.1371/journal.pone.0246640] [Medline: 33544775]
21. Chhabra G, Vashisht V, Ranjan J. A comparison of multiple imputation methods for data with missing values. Indian J Sci Technol 2017;10(19):1-7. [doi: 10.17485/ijst/2017/v10i19/110646]
22. Lo AW, Siah KW, Wong CH. Machine-learning models for predicting drug approvals and clinical-phase transitions. SSRN Journal 2017:1-60. [doi: 10.2139/ssrn.2973611]
23. Thomas T, Rajabi E. A systematic review of machine learning-based missing value imputation techniques. Data Technol Appl 2021 Apr 02;55(4):558-585. [doi: 10.1108/dta-12-2020-0298]
24. Jadhav A, Pramod D, Ramanathan K. Comparison of performance of data imputation methods for numeric dataset. Appl Artif Intell 2019 Jul 04;33(10):913-933. [doi: 10.1080/08839514.2019.1637138]
25. Enders CK. Applied Missing Data Analysis. New York, NY: Guilford Press; 2010.
26. Gimpy M. Missing value imputation in multi attribute data set. Int J Comput Sci Inf Technol 2014;5(4):1-7.
27. Kaiser J. Dealing with missing values in data. J Syst Integr 2014;5(1):42-51. [doi: 10.20470/jsi.v5i1.178]
28. Al-Helali B, Chen Q, Xue B, Zhang M. A hybrid GP-KNN imputation for symbolic regression with missing values. 2018 Presented at: Australasian Joint Conference on Artificial Intelligence; December 2018; Wellington, New Zealand p. 345-357. [doi: 10.1007/978-3-030-03991-2_33]
29. Al-Janabi S, Alkaim AF. A nifty collaborative analysis to predicting a novel tool (DRFLLS) for missing values estimation. Soft Comput 2019 Apr 11;24(1):555-569. [doi: 10.1007/s00500-019-03972-x]
30. Miyakoshi Y, Kato S. A missing value imputation method using a Bayesian network with weighted learning. Electron Commun Jpn 2012 Nov 22;95(12):1-9. [doi: 10.1002/ecj.11449]
31. Steyerberg EW. Dealing with missing values. In: Clinical Prediction Models. New York, NY: Springer; 2009:115-137.
32. Austin PC, White IR, Lee DS, van Buuren S. Missing data in clinical research: a tutorial on multiple imputation. Can J Cardiol 2021 Sep;37(9):1322-1331 [FREE Full text] [doi: 10.1016/j.cjca.2020.11.010] [Medline: 33276049]
33. Kokla M, Virtanen J, Kolehmainen M, Paananen J, Hanhineva K. Random forest-based imputation outperforms other methods for imputing LC-MS metabolomics data: a comparative study. BMC Bioinformatics 2019 Oct 11;20(1):492-502 [FREE Full text] [doi: 10.1186/s12859-019-3110-0] [Medline: 31601178]
34. Abdelkhalek I, Ben Brahim A, Essousi N. A new way of handling missing data in multi-source classification based on adaptive imputation. 2018 Presented at: 8th International Conference on Model and Data Engineering; October 24-26, 2018; Marrakesh, Morocco p. 125-136. [doi: 10.1007/978-3-030-00856-7_8]
35. Tiwari P, Colborn KL, Smith DE, Xing F, Ghosh D, Rosenberg MA. Assessment of a machine learning model applied to harmonized electronic health record data for the prediction of incident atrial fibrillation. JAMA Netw Open 2020 Jan 03;3(1):36-47 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.19396] [Medline: 31951272]
36. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002 Jun 01;16:321-357. [doi: 10.1613/jair.953]
37. Segura-Bedmar I, Colón-Ruíz C, Tejedor-Alonso M, Moro-Moro M. Predicting of anaphylaxis in big data EMR by exploring machine learning approaches. J Biomed Inform 2018 Nov;87:50-59 [FREE Full text] [doi: 10.1016/j.jbi.2018.09.012] [Medline: 30266231]
38. Lin W, Tsai C, Hu Y, Jhang J. Clustering-based undersampling in class-imbalanced data. Inf Sci 2017 Oct;409-410:17-26. [doi: 10.1016/j.ins.2017.05.008]
39. Ringnér M. What is principal component analysis? Nat Biotechnol 2008 Mar;26(3):303-304. [doi: 10.1038/nbt0308-303] [Medline: 18327243]
40. Firth D. Bias reduction of maximum likelihood estimates. Biometrika 1993;80(1):27-38. [doi: 10.1093/biomet/80.1.27]
41. Varshavsky R, Gottlieb A, Linial M, Horn D. Novel unsupervised feature filtering of biological data. Bioinformatics 2006 Jul 15;22(14):507-513. [doi: 10.1093/bioinformatics/btl214] [Medline: 16873514]
42. Anowar F, Sadaoui S, Selim B. Conceptual and empirical comparison of dimensionality reduction algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, t-SNE). Comput Sci Rev 2021 May;40:100-110 [FREE Full text] [doi: 10.1016/j.cosrev.2021.100378]
43. Banerjee M, Pal NR. Feature selection with SVD entropy: some modification and extension. Inf Sci 2014 Apr;264:118-134. [doi: 10.1016/j.ins.2013.12.029]
44. Kalankesh L, Weatherall J, Ba-Dhfari T, Buchan I, Brass A. Taming EHR data: using semantic similarity to reduce dimensionality. Stud Health Technol Informno 2013;192(1-2):52-56. [doi: 10.3233/978-1-61499-289-9-52]
45. Lou J, Cao Y, Yu Y, Hu L, Mao Z, Huang P, et al. Investigation of heart lipid changes in acute β-AR activation-induced sudden cardiac death by time-of-flight secondary ion mass spectrometry. Analyst 2020 Aug 24;145(17):5889-5896. [doi: 10.1039/d0an00768d] [Medline: 32662451]
46. Huang T, Li J, Zhang W. Application of principal component analysis and logistic regression model in lupus nephritis patients with clinical hypothyroidism. BMC Med Res Methodol 2020 May 01;20(1):99-110 [FREE Full text] [doi: 10.1186/s12874-020-00989-x] [Medline: 32357838]

47. Emergency Data Processing. 2022 Sep 20. URL: https://github.com/jacksoncoki/Emergency-data-processing [accessed 2022-09-20]
48. Steyerberg EW. Validation of prediction models. In: Clinical Prediction Models. New York, NY: Springer; 2019:329-344.
49. Geeven G, van Kesteren RE, Smit AB, de Gunst MCM. Identification of context-specific gene regulatory networks with GEMULA--gene expression modeling using LAsso. Bioinformatics 2012 Jan 15;28(2):214-221. [doi: 10.1093/bioinformatics/btr641] [Medline: 22106333]
50. Breiman L. Random Forests--Random Features: Technical Report 567. 1999 Sep. URL: https://www.stat.berkeley.edu/~breiman/random-forests.pdf [accessed 2022-12-12]
51. Ridgeway G. Generalized boosted models: a guide to the gbm package. Update 2007 Aug 3;1(1):2007.
52. Jakkula V. Tutorial on Support Vector Machine. Tutorial on Support Vector Machine (SVM) 2006:100-111 [FREE Full text] [doi: 10.1007/springerreference_106815]
53. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc: B (Methodol) 2018 Dec 05;58(1):267-288. [doi: 10.1111/j.2517-6161.1996.tb02080.x]
54. Yu Z, Wongvibulsin S, Daya NR, Zhou L, Matsushita K, Natarajan P, et al. Machine learning for sudden cardiac death prediction in the atherosclerosis risk in communities study. medRxiv. Preprint posted online January 16, 2022. [doi: 10.1101/2022.01.12.22269174]
55. Abdi H, Williams LJ. Principal component analysis. WIREs Comp Stat 2010 Jun 30;2(4):433-459. [doi: 10.1002/wics.101]
56. Villagrana-Bañuelos KE, Galván-Tejada CE, Galván-Tejada JI, Gamboa-Rosales H, Celaya-Padilla JM, Soto-Murillo MA, et al. Machine learning model based on lipidomic profile information to predict sudden infant death syndrome. Healthcare (Basel) 2022 Jul 14;10(7):1303-1318 [FREE Full text] [doi: 10.3390/healthcare10071303] [Medline: 35885829]
57. Bhattacharya M, Lu D, Kudchadkar SM, Greenland GV, Lingamaneni P, Corona-Villalobos CP, et al. Identifying ventricular arrhythmias and their predictors by applying machine learning methods to electronic health records in patients with hypertrophic cardiomyopathy (HCM-VAr-risk model). Am J Cardiol 2019 May 15;123(10):1681-1689 [FREE Full text] [doi: 10.1016/j.amjcard.2019.02.022] [Medline: 30952382]
58. Alba AC, Agoritsas T, Walsh M, Hanna S, Iorio A, Devereaux PJ, et al. Discrimination and calibration of clinical prediction models: users' guides to the medical literature. JAMA 2017 Oct 10;318(14):1377-1384. [doi: 10.1001/jama.2017.12126] [Medline: 29049590]
59. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Br J Surg 2015 Feb;102(3):148-158 [FREE Full text] [doi: 10.1002/bjs.9736] [Medline: 25627261]
60. Robertson T, Wright FT. Consistency in generalized isotonic regression. Ann Stat 1975 Mar 1;3(2):350-362. [doi: 10.1214/aos/1176343061]

Abbreviations

AUC: area under the curve
AUPRC: area under the precision-recall curve
AUROC: area under the receiver operating characteristic curve
ED: emergency department
EMR: electronic medical record
EWS: early real-time early warning system
GBM: gradient boosting machine
LASSO: least absolute shrinkage and selection operator
LDA: linear discriminant analysis
MAR: missing at random deletion
MCAR: missing completely at random
ML: machine learning
MNAR: not missing at random
LR: logistic regression
OR: odds ratio
PC: principal component
PCA: principal component analysis
PR: precision-recall
RF: random forest
ROC: receiver operating characteristic
SCD: sudden cardiac death
SVM: support vector machine

Edited by C Lovis, J Hefner; submitted 12.04.22; peer-reviewed by L Min, B Puladi; comments to author 26.07.22; revised version received 20.09.22; accepted 06.12.22; published 20.01.23.

Please cite as:
Chen X, Chen H, Nan S, Kong X, Duan H, Zhu H
Dealing With Missing, Imbalanced, and Sparse Features During the Development of a Prediction Model for Sudden Death Using Emergency Medicine Data: Machine Learning Approach
JMIR Med Inform 2023;11:e38590
URL: https://medinform.jmir.org/2023/1/e38590
doi:10.2196/38590
PMID:36662548

Original Paper

Barriers and Opportunities for the Use of Digital Tools in Medicines Optimization Across the Interfaces of Care: Stakeholder Interviews in the United Kingdom

Clare Tolley[1,2], PGCert, MPharm, PhD; Helen Seymour[3], BPharm, PGDip; Neil Watson[1,2], MSc, MBA; Hamde Nazar[1], MPharm, PhD; Jude Heed[1], BPharm, MSc; Dave Belshaw[3], RMN, MPA

[1]School of Pharmacy, Newcastle University, Newcastle upon Tyne, United Kingdom

[2]Pharmacy Department, The Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne, United Kingdom

[3]North East and North Cumbria Academic Health Science Network, Newcastle upon Tyne, United Kingdom

Corresponding Author:
Clare Tolley, PGCert, MPharm, PhD
Pharmacy Department
The Newcastle upon Tyne Hospitals NHS Foundation Trust
Queen Victoria Road,
Newcastle upon Tyne, NE1 4LP
United Kingdom
Phone: 44 0191 282 4488
Email: clare.brown@newcastle.ac.uk

Abstract

Background: People with long-term conditions frequently transition between care settings that require information about a patient's medicines to be transferred or translated between systems. This process is currently error prone and associated with unintentional changes to medications and miscommunication, which can lead to serious patient consequences. One study estimated that approximately 250,000 serious medication errors occur in England when a patient transitions from hospital to home. Digital tools can equip health care professionals with the right information at the right time and place to support practice.

Objective: This study aimed to answer the following questions: what systems are being used to transfer information across interfaces of care within a region of England? and what are the challenges and potential opportunities for more effective cross-sector working to support medicines optimization?

Methods: A team of researchers at Newcastle University conducted a qualitative study by performing in-depth semistructured interviews with 23 key stakeholders in medicines optimization and IT between January and March 2022. The interviews lasted for approximately 1 hour. The interviews and field notes were transcribed and analyzed using the framework approach. The themes were discussed, refined, and applied systematically to the data set. Member checking was also performed.

Results: This study revealed themes and subthemes pertaining to 3 key areas: transfer of care issues, challenges of digital tools, and future hopes and opportunities. We identified a major complexity in terms of the number of different medicine management systems used throughout the region. There were also important challenges owing to incomplete patient records. We also highlighted the barriers related to using multiple systems and their subsequent impact on user workflow, a lack of interoperability between systems, gaps in the availability of digital data, and poor IT and change management. Finally, participants described their hopes and opportunities for the future provision of medicines optimization services, and there was a clear need for a patient-centered consolidated integrated health record for use by all health and care professionals across different sectors, bridging those working in primary, secondary, and social care.

Conclusions: The effectiveness and utility of shared records depend on the data within; therefore, health care and digital leaders must support and strongly encourage the adoption of established and approved digital information standards. Specific priorities regarding understanding of the vision for pharmacy services and supporting this with appropriate funding arrangements and strategic planning of the workforce were also described. In addition, the following were identified as key enablers to harness the benefits of digital tools to support future medicines optimization: development of minimal system requirements; enhanced IT system management to reduce unnecessary repetition; and importantly, meaningful and continued collaboration with clinical and IT stakeholders to optimize systems and share good practices across care sectors.

(JMIR Med Inform 2023;11:e42458) doi:10.2196/42458

KEYWORDS

health information exchange; patient safety; medicines optimization; transfer of care; health informatics; qualitative

Introduction

Background

Medicines are the most common therapeutic intervention in the United Kingdom's National Health Service (NHS). However, it is estimated that between 30% and 50% of medicines prescribed for long-term conditions are not taken as intended [1]. More than a quarter of the adult population in England live with ≥2 conditions [2], and approximately 15% of people in England take ≥5 medicines a day, with 7% taking ≥8 medicines per day [3]. Furthermore, the prevalence of multimorbidity is estimated to increase substantially, with the proportion of those with >4 diseases almost doubling to 17% by 2035 [4]. Therefore, it is important that individuals receive the maximum benefit from their medicines while minimizing harm. Effective medicines optimization contributes to improved health outcomes, patient care, safety and satisfaction, improved efficiency and use of resources, better use of professional skills, and effective clinical governance [5,6].

Transfer of Care

People with long-term health conditions frequently transition between care settings; consequently, information about a patient's medicines is regularly transferred or translated between systems. However, the point at which patients transfer across different interfaces of care is high risk and is associated with unintentional changes to medications, errors, and miscommunication [7]. This can have consequences for patients, health care professionals, and the health system as a whole [8]. For example, each interaction with a health care professional may result in medication or treatment changes. Problematic polypharmacy may then occur, whereby multiple medicines are prescribed inappropriately or in which the intended benefits of the prescribed medications are not realized [3].

The Department of Health and Social Care's report, *Good for you, good for us, good for everybody*, highlighted that to reduce overprescribing, there is a need for better shared decision-making with patients; better guidance and support for clinicians; more alternatives to medicines, such as physical and social activities and talking therapies; and more structured medication reviews for those with long-term health conditions [3]. However, these goals and initiatives must be supported by effective digital systems that are interoperable and must equip health care professionals with the right information to optimize a patient's medications [9]. A systematic review of 13 publications found that the use of IT applications such as electronic health records (EHRs), electronic decision support tools, and electronic communication applications had a positive impact on financial and health outcomes [10]. However, contemporaneous and accurate information is often not available, with substantial local variation in practice, which can result in increased workload, duplication of tasks, and errors [11,12]. Currently, there is a complex network of different systems that contain patient health record data in distinct silos throughout a patient's journey. There is also a range of services available to support medicines optimization activities in the United Kingdom, including those that target transitions in care. For example, the Discharge Medicines Service, New Medicines Service, and NHS Community Pharmacist Consultation Service [6,13,14]. Plans also exist to roll out the electronic prescription service to secondary care and other care settings and to develop a patient-centered consolidated medication record that can be used by health care professionals working in different settings [15]. Such records may be associated with a range of benefits including improved safety, greater flexibility, enhanced ability to respond to patient queries, reduced duplication, and lower costs [16-18]. However, their use is still at an early stage, and challenges have been identified in some studies, including problems with system reliability, technical issues, and patient concerns regarding inaccuracies and the governance around sharing data [17,19]. The government recognizes the need for information to be collected once and then shared among providers to meet an individual's needs. Interoperability is defined as "the ability of two or more systems or components to exchange information and to use the information that has been exchanged" [20]. Nationally, a range of work is underway to enhance interoperability within health and social care settings and is a clear priority for the United Kingdom [21]. A recent policy paper, *Data saves lives: reshaping health and social care with data*, highlights commitments to introducing clear and open standards to make it easier to share data safely and efficiently across care settings [22].

Objectives

When considering how services might be delivered in the future and the necessary digital transformation, it is important to understand the current landscape of systems including their benefits, challenges, and opportunities. In this study, we aimed to address the following questions: *what systems are being used to transfer information across interfaces of care? and what are the challenges and potential opportunities for more effective cross-sector working to support medicines optimization?*

Methods

Overview

The aim of this study was to engage with key stakeholders in medicines optimization and IT across the North East and North Cumbria (NENC) integrated care system (ICS) to scope out current systems related to the transfer of information across interfaces of care. ICSs were established across regions of England on July 1, 2022, and have been described as *partnerships of organizations that come together to plan and deliver joined up health and care services.* Digital solutions will be central to supporting the function and role of ICSs. In addition, we sought to identify the challenges and potential

opportunities for a more effective cross-sector working to support medicines optimization and inform future priorities.

A qualitative methodology was selected to ensure gathering of a detailed understanding of participants' experiences and perspectives. A constructivist and interpretivist approach was taken [23] together with the framework approach, which is a method developed for use in applied policy research in which there is a need to address a clear set of aims and objectives, while following an inductive approach that allows theories to develop "bottom-up" [24].

Key stakeholders in medicines optimization and digital health were invited to participate in a semistructured interview to gather their perspectives on the current medicines optimization services provided and the digital tools used to support these activities. In addition, stakeholders were asked to provide their opinions on the barriers and potential opportunities for more effective medicines optimization across sectors.

Eligibility Criteria

We included all clinicians, managers, commissioners, and stakeholders who were or had been involved in commissioning, developing, and delivering medicines optimization–related activities and who had expertise in IT, clinical informatics, and digital health solutions used within the NHS.

Recruitment

We used a snowball sampling approach to identify suitable key participants [25]. In the first instance, the researcher was introduced to an initial set of contacts by the senior medicines optimization pharmacist based at the NENC Academic Health Science Network and the digital transformation director for the NENC Academic Health Science Network. The researcher emailed potential participants and invited them to participate in a semistructured interview. This email also included attachments to a participant information leaflet and a consent form. The participants were required to provide consent via a web-based consent form before participating in the study. We proactively aimed to engage with individuals from a range of professional backgrounds and levels of experience to ensure that the data gathered were rich and representative. Data were collected until thematic saturation was reached, and we used an inductive approach to look for the nonemergence of new themes as interviews and analysis were conducted [25,26].

Data Collection

Semistructured interviews were conducted by 1 researcher (CT) between January and March 2022 to explore stakeholders' perspectives on the current medicines optimization services provided in the region and the digital tools used to support these activities, including details regarding the flow of health information exchange and interoperability reached. In addition, stakeholders were asked to provide their opinions on the barriers and potential opportunities for more effective medicines optimization across sectors. A flexible topic guide was developed that incorporated open-ended questions and prompts. The guide was shared with a team of researchers and clinicians to review and refine the data before use, and the guide was developed iteratively throughout data collection [27]. Interviews lasted for approximately 1 hour and were conducted via video call by a researcher with clinical and postgraduate level of qualitative data collection experience, at a mutually convenient time for each participant. All interviews were recorded, transcribed verbatim together with accompanying field notes, and anonymized.

Analysis (Interviews)

Qualitative data collection and analysis were iterative, allowing themes to be generated, interpreted, explored, and disconfirming evidence identified [28]. Different data sources, for example, interviews with a range of participants, facilitated triangulation to identify where and how different data converged and diverged. The main themes and subthemes were identified using a constant comparative analysis [28]. For this purpose, data were constantly compared among interviewees to explore similarities and differences between groups and to uncover explanations for why these differences existed. Field notes contributed to the analysis by providing valuable context, for example, in which the participants used their voice to stress points or in which humor was used. Field notes were also used by the interviewer to note their own reflections and consider questions for future exploration [29]. The framework approach was used, which is a 5-staged approach to thematic analysis, enabling previous theories and insights identified through literature review or experience, to inform the development of the thematic framework, while allowing theme generation based on the data and was, therefore, open to discovering unexpected concepts based on the participants' experiences. This was used as a complementary method along with the constant comparative analysis. Themes were discussed among team members and continually refined and applied systematically to the whole data set using the computerized software N-Vivo (QSR International). All data were analyzed by qualified members of the research staff. "Member checking" was also performed, whereby a draft of the key findings was shared with all participants, who were given a minimum of 2 weeks to provide feedback on the interpretations made and contribute any additional insight [26].

Ethics Approval

This study was approved by the Research, Policy, Intelligence, and Ethics team at Newcastle University (reference: 17851-2021).

Results

Overview

A total of 22 interviews were conducted with 23 participants lasting between 38 and 75 minutes (Table 1). Two participants (a community pharmacist and general practitioner [GP]) were unable to take part in an interview, owing to clinical commitments and availability. "Member checking" resulted in 1 correction to the results (clarification regarding work underway to create a patient medication records) and provided further information regarding national initiatives currently underway to support development of shared care records, which was incorporated into the discussion and recommendations in the *Addressing Digital Gaps* section.

This study revealed a range of different systems that are used across a region to support medicines optimization activities. From the interviews, a range of themes and subthemes pertaining to three key areas were identified: (1) transfer of care issues, (2) challenges of digital tools, and (3) future hopes and opportunities.

Table 1. Table of participants (N=23).

Profession and Sector	Participants, n (%)
Pharmacists (n=21)	
Community	4 (17)
Hospital	5 (22)
Primary care (GP[a] practice, primary care network, and CCG[b])	5 (22)
North East Ambulance Service	2[c] (9)
NHS[d] England and NHS Improvement	5 (22)
GPs	2 (9)

[a]GP: general practitioner.

[b]CCG: clinical commissioning group.

[c]One interview conducted with both participants.

[d]NHS: National Health Service.

Medicines Optimization Systems

Informed by participant interviews and relevant literature, a simplified overview of the key systems related to medicines across primary, secondary, tertiary, and social care across the NENC was developed (Multimedia Appendix 1 [30-32]). This was not intended to provide a comprehensive overview of all systems used throughout the region but instead to illustrate the complexity of how data are stored and moved between settings. Several patient health record systems have been used in general practice and primary care. A range of different community pharmacy patient medication record systems was also identified, which lacked the ability to directly transfer information with GP systems. At the time of data collection, some hospital trusts in the region used paper-based health records and prescriptions. The transfer of information between different systems and care settings was largely facilitated through bespoke solutions delivered by third-party companies in response to a particular problem, for example, a digital referral as part of the Discharge Medicines Service from 1 hospital to a community pharmacy was typically either sent using NHS mail or a website or integrated web platform such as PharmOutcomes or Cegedim. The information flow between the systems was found to be typically unidirectional.

Transfer of Care Issues

All participants highlighted problems owing to incomplete patient records. Data were described as being held in separate silos by the GP, secondary care providers, or a community pharmacy, with ineffective data flow among them, resulting in reduced efficiency and safety. For example, participants described discrepancies between the allergy status recorded in different clinical systems (eg, missing documented allergies) and omissions owing to poor communication and clinical handover. Incomplete records also made it difficult to proactively provide care by identifying patterns of behavior that would warrant further investigation or management, for example, “if someone’s getting emergency contraception on a regular basis [from a community pharmacy], actually that should be flagging up a risk” (GP 006). A GP recalled a “significant incident where the GP hadn’t put down [methotrexate] or- on the GP record was not methotrexate, and they [the patient] ended up in an ITU in Wales and [the staff] didn’t know that the patient was on methotrexate and had actually accidentally overdosed” (GP 006), but the staff were unaware that the patient was even prescribed this medication.

Challenges of Digital Tools

Participants raised concerns on specific challenges associated with the digital tools used to optimize medications. These included the use of multiple systems and workflow, interoperability, digital gaps, IT systems management, and change management.

Multiple Systems and Workflow

Participants working across all settings described the need to interact with multiple IT systems as a part of their day-to-day role. Those working in general practice, for example, would access a core EHR system alongside other systems for viewing additional information, such as hospital notes, appointment letters, or blood tests. Community pharmacists described navigating between a growing number of different systems to fulfill different tasks and purposes (refer to the quote below) and felt that exposure to “more interfaces, [presented] [...] more opportunities [...] for information to have to be re-transcribed [and] [...] actually alert tasks to get dropped, because the right people can’t see [what needs to be done]” (Pharmacist 010). They emphasized how “the more you’re having to step out of your day-to-day workflows and go, ‘Oh my goodness, I really must remember to send a DMS [Discharge Medicines Service] referral for this thing’. You’re just not going to do it” (Pharmacist 010):

> *On a Sunday, I do a COVID clinic, so I've got Q-Flow open, which is the appointments booking system [...] I’ve got PharmOutcomes open for CPCS referrals*

> *via 111. I've got PharmOutcomes open for other things, any other bits and pieces that might come. I've got Outcomes4Health, which is the sister platform. That's open for recording the COVID vaccinations. Then we've got the...What else is there? There are half a dozen different things there. Of course, in the pharmacy clinical system, we [they] use Positive Solutions Analyst, but we'd have that open as well.* [Pharmacist 007]

Interoperability and Safety

A GP described how community nurse practitioners do not have access to the full EHR of certain patients in their care. This occurred because the community nurses were employed by the local hospital rather than by the GP practice, and the 2 organizations used different IT systems, which did not share data in real time. Consequently, automated checks such as drug-laboratory checks or drug-interaction checks are not reliable. There was the risk that if community nurses "don't have the blood results in their system [so] they will not get that [computerised] warning, so they may merrily go ahead and prescribe that [medication]. Then conversely, if they've prescribed a drug, I [the GP] don't know that the interactions are there now" (GP 012).

Digital Gaps

The participants revealed important gaps in the availability of health data that could not be easily shared or used by health care professionals, for example, paper-based hospital notes. A locally shared health record for people living in NENC known as the "Great North Care Record" (GNCR) is currently in development, with the aim of providing health and care workers access to current medical information. However, 1 GP highlighted how their local acute hospital trust only imported some "very, very, primitive data" into the GNCR and so was not "hugely valuable" (GP 012). In addition, data from social care, for example, care homes, were not imported into the GNCR, which meant that "any changes that are made there are not displayed" (GP 006). In contrast, however, more digitally advanced hospitals transferred a greater quantity of information from their EHR into the GNCR, which was "more useful" (GP 012). This raised the possibility of inequalities among health care providers because organizations "that are [digitally] further behind are the ones that...The Great North Care Record doesn't help" (GP 012).

Some professional sectors, such as community pharmacy, were also unable to access the GNCR, which made it "really difficult" to maximize use of that workforce and "shift patients away from some of those higher acuity services" (Pharmacist 003). Furthermore, information governance arrangements for accessing multiple systems were considered problematic. For example, 1 pharmacist suggested that the current consent model whereby community pharmacists must obtain consent before accessing information within the summary care record was inappropriate because "if you've got a patient's prescription it should almost imply by informed consent, they've given you their script" (Pharmacist 005).

Participants also revealed a low adoption of standardized digital codes for documenting clinical data within health systems. There are "a lot of trusts [that]don't use dm+d" (Pharmacist 002), and many hospitals are "not using SNOMED at the moment" (GP 006); without adoption of the standards, large initiatives such as shared patient medication records are "not going to work" (Pharmacist 002). Dictionary of medicines and devices is a dictionary of descriptions and codes, which represent medicines and devices used across the NHS. Systematized Nomenclature of Medicine Clinical Terms is a structured clinical vocabulary for use in EHRs and covers diagnoses, procedures, etc.

Finally, the participants noted how certain services were being rolled out without a supportive digital infrastructure. For example, the national hypertension case finding service in community pharmacies lacked "a national system for reporting it [blood pressure] [and reporting] between the two [pharmacy to GP]" (Pharmacist 011). Consequently, a range of different communication techniques, including email, letter, or pilot digital platforms, were being used, depending on local arrangements, and as data were not collected in a standardized digital format, it could not be transferred between care settings in an interoperable way.

IT Systems Management

Participants suggested that a lack of minimum standards or mandating how services are technically delivered is problematic and contributes to low adoption and delays in rolling out clinical services and digital solutions:

> *So pharmacy DMS transmissions, we launched the DMS service but pharmacies [are] not seeing many of them because yes you can do it by snail mail and NHS Mail, but that's not good enough. We shouldn't launch a service without a platform to deliver it on. So every hospital should be told, "You can have your own, but we're launching this new service and you must be able to provide a digital solution that's integrated."* [Pharmacist 005]

A need was identified to rationalize the number of digital systems used, create clear expectations for suppliers, and develop standards and frameworks that outline how services should be digitally enabled. This would reduce waste from "reinventing the wheel every single time" (Pharmacist 020). A new service or digital tool was developed to limit duplication and unnecessary costs. Another pharmacist added the following:

> *I think there needs to be a suite of expectations that everybody needs to have and it's the same with hospitals, you can't have a hospital system that doesn't do these 10 things [...] every system needs to do that by a certain deadline.* [Pharmacist 005]

Change Management

Change management was seen as important for managing the future development of digital medicines optimization services. For instance, although the clinical terminology standards have now been defined in the Systematized Nomenclature of Medicine (SNOMED) and dictionary of medicines and devices (dm+d), "the hard bit is to come, which is the adoption and more importantly the transformation around that. There's also

a Hearts and Minds piece" (Pharmacist 002). Similarly, participants described concerns regarding data sharing, for instance:

> *If you suggest [data sharing] to some general practices, they'll say: "No chance. Nobody should have access to that data." But once you get over that barrier, and there's a bit of trust built into it, then you can start adding to it.* [Pharmacist 003]

Instead, all health care professionals need to work together to "best serve the needs of the patients" (Pharmacist 019).

Future Hopes and Opportunities

Consolidated Integrated Care Record

Participants revealed several ambitions for the future of digital medicines optimization services, although the need for a single shared consolidated medication record, giving all health and care professionals access to data, was a clear priority and reflected those working across care sectors:

> *One record across all organizations, that's the blue-sky thinking. The data would sit in a data repository that would be coded and accessible, via APIs, via front-end systems that could be customised to be targeted to how GPs work, targeted too how secondary care clinicians work, targeted for acute and outpatient mental health. Fundamentally, all the data would be held in one central repository for that patient and all of the systems can pull in all that data. A patient, for example from a prescribing basis, would have one prescribing record. That prescribing record would continue out of hospital, into hospital.* [GP 012]

This would "open up a number of opportunities [...] for optimisation and proper management of patients and overprescribing" (Pharmacist 002). Furthermore, by identifying and using common digital architectures and standards, such records may be linked in the future.

Some participants were worried that increased access to information could be problematic in some situations, for example, community pharmacists could have "too much information, to make a decision on" because "you could spend hours and hours and hours trailing back through communications and stuff that gets put onto records, [which] it's probably completely inappropriate to the query" (Pharmacist 018). Consequently, several participants suggested involving end users in the development of shared records and posed that for "a sector and a workforce like community pharmacy, I think it's big enough to warrant having a bespoke solution developed for it" (Pharmacist 001). A change in funding arrangements was also discussed as vital to support and incentivize community pharmacists to deliver clinical services at scale and to justify the need for access to shared care records in the first instance.

Uses of Data

There were hopes that the enhanced availability of medication data could improve workflows across care settings and population health management. Improved efficiency may be realized, for example, by sending a coded list of a patient's medications from one system to another, so that a clinician only has to "go click, click, click, and it populates the prescribing system" (Pharmacist 010). In addition, to support better interprofessional working and better continuity of care, participants thought it would be good if they could send a "request, where you've got specific things that you want following up" (Pharmacist 010) alongside any information and context about a patient's medication, directly between systems.

Participants also discussed how the development of a comprehensive shared patient record could serve as a "population health platform" (Pharmacist 002). It would then be possible to "start interrogating the information at a patient level, but [also] at a population level" (Pharmacist 002) and target public health challenges such as "overprescribing, opiate prescribing, valproate [prescribing in pregnancy]" (Pharmacist 002). Digital tools could also support better clinical prioritization, for example:

> *If you've got 100 Primary Care Network pharmacists delivering a structured medication review every year or so, that's 1,000 reviews a week. How do you know which 1,000 patients put in for those slots? How do you caseload? That's really important.* [Pharmacist 016]

Comprehensive patient data repositories could enable clinicians to develop robust strategies to identify patients more efficiently, to avoid "each practice pharmacist going out and trying to design their own searches" (Pharmacist 013). It was clear that there are major opportunities arising from more effective data sharing, and as 1 pharmacist remarked, there are likely innovations that health care professionals have not even started to dream about.

Discussion

Principal Findings

This research has identified several challenges and potential opportunities related to the use of digital tools for delivering medicines optimization, which are categorized under 3 key concepts: transfer of care issues, challenges of digital tools, and future hopes and opportunities. There is substantial complexity in the number of various medicines management systems used throughout the region and the challenges associated with incomplete patient records. The use of multiple systems also affects the user workflow. There was a lack of intrasystem and intersystem interoperability and important gaps in digital data in some settings (eg, social care and in some areas where hospital prescribing was paper based). Several problems related to IT systems management and change management have also been described. Participants described a clear need for a patient-centered consolidated integrated health record for use by all health and care professionals across different sectors, bridging those working in primary, secondary, and social care.

We also identified a series of recommendations relevant to health service managers, policy makers, and clinical staff, which are discussed in Table 2.

Table 2. Summary table: stakeholder recommendations and objectives.

Recommendations and specific objectives	Key stakeholder group
Identify the future vision for pharmacy services and support with appropriate funding plan	
Support realization of the vision through appropriate strategic planning and funding arrangements, for example, community pharmacy contractual framework	Policy makers and service managers
Development of a patient-centered consolidated integrated health record for use by health and care professionals	
Support digitization of social care, for example, implementation of electronic prescribing and medication administration in care homes	Researchers and service managers
Creation of local multistakeholder working groups to ensure cocreation of digital solutions and facilitate effective teamwork and "buy-in"	Service managers
In-depth exploration of concerns around data sharing and governance considerations with key clinical and IT stakeholders and patients and members of the public	Policy makers, service managers, researchers, clinical staff, and public
Explore need for tailored views for health professional groups to shared care records	Service managers and researchers
Support adoption of digital solutions across NENC[a] region's hospitals with a pledge to end paper records and prescribing	Policy makers, service managers, and clinical staff
Adoption of medication standards across NHS[b] (ie, dm+d[c] and SNOMED CT[d])	
Identify what is the level of dm+d adoption across the region	Policy makers and service managers (this is in progress via NHS Digital)
Explore the local facilitators and barriers to adoption of standards	Policy makers, service managers, and researchers
Prioritize communication around the need to adopt standards locally	Policy makers, service managers, and clinical staff
Support sites through the transformation process, using tools from NHS England or FCI[e]	Service managers
Rationalizing number of systems and services to reduce unnecessary repetition	
Mapping process to identify services provided across sectors and highlight duplication and gaps	Service managers and clinical staff
Identify clinical services that would benefit from digitization in the community pharmacy sector and to support service provision and service management or audit	Service managers and researchers
Development of frameworks and minimum standards outlining how services should be delivered using digital means at a local and national level	
Monitor guidance and toolkits from NHS organizations (eg, NHS Transformation Directorate) and publish updates, publish bulletins, and organize workshops to increase awareness of an organization's responsibilities and share lessons across the region	Policy makers, service managers, researchers, and clinical staff
Use and promote the use of forums for communication across all sectors and levels from manager to frontline staff	Service managers and clinical staff
Cross-sector communication around approaches to digital medicines optimization	
Development or use of established forums and groups to proactively collate and communicate examples of good practices between different clinical and IT stakeholders	Policy makers, service managers, and researchers
Working with suppliers to develop integrated solutions to avoid unnecessary development of bespoke solutions	
Share lessons across organizations about successful collaborations with suppliers on innovation projects	Service managers, researchers, and clinical staff
Harness insight from existing projects to support scale-up of innovations	Service managers and researchers

[a]NENC: North East and North Cumbria.

[b]NHS: National Health Service.

[c]dm+d: dictionary of medicines and devices.

[d]SNOMED CT: Systematized Nomenclature of Medicine Clinical Terms.

[e]FCI: Faculty Clinical Informatics.

Addressing Digital Gaps

The provision of health services in the United Kingdom needs to substantially change to meet the needs of an aging population, with a key focus on more integrated care across health and care settings to local needs [33]. This requires better collaboration between different professionals working across care sectors who have the right information to inform decision-making, with

digital developments recognized as key to the transformation [33,34]. To enable this, participants emphasized the need for a single consolidated electronic patient record where all health and care professionals can read and write into and share information across traditional boundaries to deliver acute care, manage long-term conditions, and ensure patients receive the right care at the right time and place. For example, most community pharmacists in the United Kingdom do not have access to an up-to-date and comprehensive list of a patient's medications or medical records, which hinders their ability to support individuals. However, pilot projects such as the Somerset shared care record case study (SIDeR) are underway with some promising feedback, although there is a need for further evaluation to fully understand the benefits and any unintended consequences or challenges [35]. There is a strong relationship between the maturity of digital health and comprehensive evaluation methods; therefore, this should be included as part of future locally shared care strategies to support learning and facilitate the development and adoption of systems [36]. In addition, as Cresswell et al [37] noted, there are several socio-organizational dimensions of change that must be considered to support a digitally enabled shared care agenda. They summarize key areas, such as structural and organizational complexity; variations in data management and expectations; poorly defined current shared care pathways; and issues associated with reluctance to data sharing, managing "data overload," and configuring systems appropriately. In terms of technological dimensions, existing infrastructures and legacy systems may hinder data sharing across new technological junctions, while it may also be difficult to connect incompatible data structures and overcome supplier resistance to make distinct systems interoperable [37]. In their report, they advocated the need to map potential architectural components and designs for shared care solutions with careful consideration of their potential benefits and limitations [37].

In our study, participants highlighted that the need to undertake the process of developing shared records *with* end users to ensure the design and functionality is "fit for purpose." Specifically, participants raised the question of developing tailored solutions, that is, a bespoke community pharmacy view of a patient's health record, to enhance the usability and utility of such systems. Research has shown how the design of EHRs can influence behavior and prescribing safety [38,39]; therefore, further work is needed to explore how to create usable health records for a range of different end users working across sectors.

Participants also expressed specific ways in which data could be used to better support the transfer of care, for instance, sending digital referrals or requests for follow-up directly between existing clinical systems, further demonstrating the need for continued engagement and collaboration with end users involved in the delivery of frontline services to optimize and enhance systems over time [40].

To fully harness the benefits of a comprehensive and consolidated shared patient record, there is a need to address digital gaps within organizations across the health and social care sector. The effectiveness and utility of the tool depends on the data within. First, organizations must prioritize the use of established and approved digital information standards related to medicines and clinical information within NHS digital systems. Notably, all NHS care providers who are involved in prescribing, dispensing, or administering medicines must transfer medication information using the newest UK version of fast health care interoperability resources, use approved dose syntax to transfer the amount of medication per dose as a simple coded quantity, and use SNOMED and dm+d codes for allergy or intolerance information by March 31, 2023 [41]. However, barriers to the adoption of such standards have been identified, including a lack of cohesive national-scale digital health system; funding and support for standards; knowledge and infrastructure related, such as the impact on preexisting workflows; and lack of use of a consistent patient ID [42,43]. Therefore, local ICSs must explore how they can address such challenges and support the implementation of the standards across the region. This underscores the importance of effective clinical leadership and understanding the personal factors that influence health IT uptake [44,45]. The adoption of digitized health records, electronic prescriptions, and medication administration across primary, secondary, and social care is vital. As is embedding digital technologies across social care, particularly as estimates suggest that less than half of social care providers have any form of digital care records in England [46]. Furthermore, research has shown that a large proportion of medication errors occur in care homes [47]; therefore, tools to support the digitization of the sector to support medicines optimization and enhance the safety, quality, and efficiency are urgently needed. Our findings echo those of a recently published report by the Royal Pharmaceutical Society Scotland, *Pharmacy 2030: a professional vision document*, which outlines the changes to and enhanced roles of pharmacy professionals and key enablers, including data to inform decision-making, harnessing digital technology, developing the workforce, and increasing emphasis on multidisciplinary work [34]. A further report and policy review from the Royal Pharmaceutical Society and The King's Fund that will inform the development of vision for pharmacy was published at the end of 2022.

To support the delivery of enhanced medicines optimization activities across care settings, there will be an increased emphasis on pharmacists and pharmacy staff undertaking clinical roles in all sectors. Community pharmacies, for example, will have a far greater role in providing enhanced clinical services and supporting the holistic prevention of ill health in a community [48,49]. The supply of medicines will be facilitated by accuracy checking technology, such as dispensing robots, or possibly through hub-and-spoke model dispensing [50]. To support this, the participants in this study highlighted the need for changes in how services are funded through the community pharmacy contractual framework. Although the demand for community pharmacies has risen since the outbreak of the Covid-19 pandemic, staff shortages are a growing concern; 1 survey reported that 91% (estimated from responses from 418 representatives of 5000 pharmacy premises) of pharmacies have experienced staff shortages [51]. There are also huge concerns around staffing shortages more widely across the NHS, which has an impact on patient care, while awaiting the results and recommendations of the NHS long-term workforce strategy [52]. This is important because changes resulting from the implementation of new digital tools, particularly systems with

poor usability, can contribute to clinician burnout and consequently reduce job satisfaction, quality and safety of care, and costs [53]. Any digital transformation relating to how medicines optimization is delivered should, therefore, be mindful of the working environment and additional stressors present, and improving the working life of health and social care providers should be a goal and actively monitored [54]. In addition, IT systems management is needed to closely monitor the number of digital systems used across an ICS footprint and rationalize how services are delivered to reduce unnecessary duplication and streamline services. This may be supported by forums and established groups with a key objective of proactively collating and communicating examples of good practices between different clinical and IT stakeholders.

Limitations

We acknowledge some limitations of this study; semistructured interviews were conducted via web-based videocall platforms, which enabled participants across a large geographical area to participate; however, we did encounter some technical issues, which may have impacted the flow and nature of discussions. We only included 2 GPs in this study; however, 5 primary care and clinical commissioning group pharmacists who provided in-depth detail about the experiences of medicines optimization of health care professionals working within GP practices were recruited, and data were collected until thematic saturation was reached. Further work may specifically explore the challenges experienced within different sectors and enhanced by using observational data collection approaches. Furthermore, the interviewer was professionally known to a small number of participants before the study, and her experience as a practicing pharmacist and academic researcher enabled her to build rapport and relationships with the participants. Throughout data collection and analysis, notes were recorded on any personal reactions or reflections to support consideration of the intersubjective reflexivity between herself and the participants [55]. This was used during the analysis stage to provide context and allowed the researcher to honestly and critically reflect on their role in interpreting the data. It also prompted further exploration of ideas and themes during the data collection stage [55]. We collected data from a range of participants across different sectors; however, this was limited to the NENC regions and so may not be representative of other parts of the United Kingdom. Finally, although the analysis and results were discussed with the research team and member checking was performed with the participants, only 1 researcher coded the data collection transcripts, which could have further decreased the rigor and replicability of our work.

Conclusions

The findings from this qualitative study of 23 clinical and IT stakeholders identified major complexity in terms of the number of different systems used throughout the NENC region and identified several important challenges in the transfer of care issues, focusing on having access to incomplete patient records. We also highlighted important barriers related to the use of digital tools, such as multiple systems and workflow, interoperability, digital gaps, IT systems management, and change management. Finally, participants discussed their future hopes and opportunities for the provision of medicines optimization services in the future, and there was a clear need for a patient-centered consolidated integrated health record for use by health and care professionals across different sectors, which would be fundamental to delivering effective and safe patient care. Further specific priorities were around understanding the vision for pharmacy services and supporting it with appropriate funding arrangements and strategic planning of the workforce, adoption of digital information standards, and development of minimal system requirements and frameworks. In addition, IT system management to reduce unnecessary repetition and, importantly, meaningful and continued collaboration with stakeholders and system suppliers to optimize systems and share good practices across care sectors were important key enablers.

Acknowledgments

This study was supported by the Academic Health Science Network and the National Health Service England Digital Primary Care First program. The funder had no input in the study design and data collection, analysis, or interpretation. The authors would like to thank the participants for their time and valuable contributions to this work.

Authors' Contributions

All the authors were responsible for the conception and study design. CT performed data collection and liaised with all authors on the analysis. CT led the writing of this manuscript, with all authors commenting on drafts. All authors have read and approved the final manuscript for submission.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Simplified overview of digital tools used across North East and North Cumbria to support medicines optimization activities.
[DOCX File , 317 KB - medinform_v11i1e42458_app1.docx]

References

1. World Health Organization. Adherence to Long-term Therapies : Evidence for Action. Geneva: World Health Organization; 2003.
2. Multiple long-term conditions (multimorbidity): making sense of the evidence. National Institute for Health and Care Research. URL: https://evidence.nihr.ac.uk/collection/making-sense-of-the-evidence-multiple-long-term-conditions-multimorbidity/ [accessed 2023-02-16]
3. Good for you, good for us, good for everybody. Department of Health & Social Care. URL: https://tinyurl.com/27cs5ncw [accessed 2023-02-16]
4. Kingston A, Robinson L, Booth H, Knapp M, Jagger C, MODEM project. Projections of multi-morbidity in the older population in England to 2035: estimates from the Population Ageing and Care Simulation (PACSim) model. Age Ageing 2018 May 01;47(3):374-380 [FREE Full text] [doi: 10.1093/ageing/afx201] [Medline: 29370339]
5. Medicines optimisation: the safe and effective use of medicines to enable the best possible outcomes. National Institute for Health and Care Excellence. 2015. URL: https://www.nice.org.uk/guidance/ng5 [accessed 2023-02-16]
6. NHS Clinical Commissioners. The role and functions of CCG medicines optimisation teams. NHS Clinical Commissioners. URL: https://www.nhsconfed.org/system/files/2021-07/Role-and-functions-of-the-CCG-medicines-optimisation-team.pdf [accessed 2023-02-06]
7. Keeping patients safe when they transfer between care providers – getting the medicines right Final report. Royal Pharmaceutical Society. 2012. URL: https://www.rpharms.com/Portals/0/RPS%20document%20library/Open%20access/Publications/Keeping%20patients%20safe%20transfer%20of%20care%20report.pdf [accessed 2023-02-16]
8. Medication safety in transitions of care technical report. World Health Organization. URL: https://www.who.int/publications/i/item/WHO-UHC-SDS-2019.9 [accessed 2023-02-16]
9. Li E, Clarke J, Neves AL, Ashrafian H, Darzi A. Electronic health records, interoperability and patient safety in health systems of high-income countries: a systematic review protocol. BMJ Open 2021 Jul 14;11(7):e044941 [FREE Full text] [doi: 10.1136/bmjopen-2020-044941] [Medline: 34261679]
10. Kooij L, Groen WG, van Harten WH. The effectiveness of information technology-supported shared care for patients with chronic disease: a systematic review. J Med Internet Res 2017 Jun 22;19(6):e221 [FREE Full text] [doi: 10.2196/jmir.7405] [Medline: 28642218]
11. Everson J, Adler-Milstein J. Gaps in health information exchange between hospitals that treat many shared patients. J Am Med Inform Assoc 2018 Sep 01;25(9):1114-1121 [FREE Full text] [doi: 10.1093/jamia/ocy089] [Medline: 30010887]
12. Warren LR, Clarke J, Arora S, Darzi A. Improving data sharing between acute hospitals in England: an overview of health record system distribution and retrospective observational analysis of inter-hospital transitions of care. BMJ Open 2019 Dec 05;9(12):e031637 [FREE Full text] [doi: 10.1136/bmjopen-2019-031637] [Medline: 31806611]
13. Discharge medicines service. Pharmaceutical Services Negotiating Committee. URL: https://psnc.org.uk/national-pharmacy-services/essential-services/discharge-medicines-service/ [accessed 2023-02-16]
14. Transfer of care initiative. NHS Digital. URL: https://digital.nhs.uk/services/transfer-of-care-initiative [accessed 2023-02-16]
15. Interoperable medicines standards and electronic prescription service (EPS). NHS Digital. URL: https://files.digital.nhs.uk/FC/EB702D/Interoperable%20Medicines%20webinar%20-%201%20March%202021.pdf [accessed 2022-05-20]
16. Menachemi N, Rahurkar S, Harle C, Vest J. The benefits of health information exchange: an updated systematic review. J Am Med Inform Assoc 2018 Sep 01;25(9):1259-1265 [FREE Full text] [doi: 10.1093/jamia/ocy035] [Medline: 29718258]
17. Compeau DR, Terry A. Connecting medical records: an evaluation of benefits and challenges for primary care practices. J Innov Health Inform 2017 Jun 30;24(2):855 [FREE Full text] [doi: 10.14236/jhi.v24i2.855] [Medline: 28749315]
18. Krauss ZJ, Abraham M, Coby J. Clinical pharmacy services enhanced by electronic health record (EHR) access: an innovation narrative. Pharmacy (Basel) 2022 Dec 05;10(6) [FREE Full text] [doi: 10.3390/pharmacy10060170] [Medline: 36548326]
19. Sanyer O, Butler J, Fortenberry K, Webb-Allen T, Ose D. Information sharing via electronic health records in team-based care: the patient perspective. Fam Pract 2021 Jul 28;38(4):468-472. [doi: 10.1093/fampra/cmaa145] [Medline: 33684209]
20. IEEE Standard Computer Dictionary: A Compilation of IEEE Standard Computer Glossaries. Piscataway, New Jersey, United States: IEEE; Jan 18, 1991.
21. Sheikh A, Anderson M, Albala S, Casadei B, Franklin BD, Richards M, et al. Health information technology and digital innovation for national learning health and care systems. Lancet Digital Health 2021 Jun;3(6):e383-e396. [doi: 10.1016/s2589-7500(21)00005-4]
22. Data saves lives: reshaping health and social care with data. Department of Health & Social Care. 2022. URL: https://tinyurl.com/2me5etrf [accessed 2023-02-16]
23. Harrison H, Birks M, Franklin R, Mills J. Case study research: foundations and methodological orientations. Qual Social Res 2017;18(1). [doi: 10.17169/fqs-18.1.2655]
24. Gale NK, Heath G, Cameron E, Rashid S, Redwood S. Using the framework method for the analysis of qualitative data in multi-disciplinary health research. BMC Med Res Methodol 2013 Sep 18;13:117 [FREE Full text] [doi: 10.1186/1471-2288-13-117] [Medline: 24047204]
25. Malterud K, Siersma VD, Guassora AD. Sample size in qualitative interview studies: guided by information power. Qual Health Res 2016 Nov;26(13):1753-1760. [doi: 10.1177/1049732315617444] [Medline: 26613970]

26. Bryant A, Charmaz K, editors. The SAGE Handbook of Grounded Theory. Thousand Oaks, California, United States: SAGE Publications; 2010.
27. DeJonckheere M, Vaughn LM. Semistructured interviewing in primary care research: a balance of relationship and rigour. Fam Med Community Health 2019;7(2):e000057 [FREE Full text] [doi: 10.1136/fmch-2018-000057] [Medline: 32148704]
28. Pope C, Ziebland S, Mays N. Qualitative research in health care. Analysing qualitative data. BMJ 2000 Jan 08;320(7227):114-116 [FREE Full text] [doi: 10.1136/bmj.320.7227.114] [Medline: 10625273]
29. Phillippi J, Lauderdale J. A guide to field notes for qualitative research: context and conversation. Qual Health Res 2018 Feb;28(3):381-388. [doi: 10.1177/1049732317697102] [Medline: 29298584]
30. Beetroot community homepage. BeetrootCommunity. URL: https://www.beetroothealth.com/beetrootcommunity/ [accessed 2022-11-29]
31. Tamonitor. Therapy Audit. URL: https://www.therapyaudit.com/tamonitor/ [accessed 2022-11-29]
32. Chemocare. CIS Oncology. URL: https://www.cis-healthcare.com/chemocare/ [accessed 2023-02-16]
33. The NHS long term plan. BMJ. 2019 Jan 07. URL: http://www.bmj.com/lookup/pmidlookup?view=long&pmid=30617185 [accessed 2023-02-16]
34. Pharmacy 2030: a professional vision. Royal Pharmaceutical Society Scotland. URL: https://www.rpharms.com/pharmacy2030 [accessed 2023-02-16]
35. PSNC briefing 022/22: Somerset shared care record case study (SIDeR). Middlesex Pharmaceutical Group. URL: https://www.middlesexlpcs.org.uk/psnc-briefing-022-22-somerset-shared-care-record-case-study-sider/ [accessed 2023-02-16]
36. Baltaxe E, Czypionka T, Kraus M, Reiss M, Askildsen JE, Grenkovic R, et al. Digital health transformation of integrated care in Europe: overarching analysis of 17 integrated care programs. J Med Internet Res 2019 Sep 26;21(9):e14956 [FREE Full text] [doi: 10.2196/14956] [Medline: 31573914]
37. Cresswell KS, Anderson S, Mozaffar H, Elizondo A, Geiger M, Williams R. Attention to socio-organisational dimensions is key to advancing the shared care record agenda in health and social care (Preprint). Research Gate 2022 Mar:38310 (forthcoming) [FREE Full text] [doi: 10.2196/preprints.38310]
38. MacKenna B, Bacon S, Walker AJ, Curtis HJ, Croker R, Goldacre B. Impact of electronic health record interface design on unsafe prescribing of ciclosporin, tacrolimus, and diltiazem: cohort study in English national health service primary care. J Med Internet Res 2020 Oct 16;22(10):e17003 [FREE Full text] [doi: 10.2196/17003] [Medline: 33064085]
39. Brown C, Mulcaster H, Triffitt K, Sittig D, Ash J, Reygate K, et al. A systematic review of the types and causes of prescribing errors generated from using computerized provider order entry systems in primary and secondary care. J Am Med Inform Assoc 2017 Mar 01;24(2):432-440 [FREE Full text] [doi: 10.1093/jamia/ocw119] [Medline: 27582471]
40. Cresswell KM, Lee L, Mozaffar H, Williams R, Sheikh A, NIHR ePrescribing Programme Team. Sustained user engagement in health information technology: the long road from implementation to system optimization of computerized physician order entry and clinical decision support systems for prescribing in hospitals in England. Health Serv Res 2017 Oct;52(5):1928-1957 [FREE Full text] [doi: 10.1111/1475-6773.12581] [Medline: 27714800]
41. DAPB4013: medicine and allergy/intolerance data transfer. NHS Digital. URL: https://tinyurl.com/yfhap3kw [accessed 2023-02-16]
42. G7 Open standards and interoperability Final Report. Department of Health & Social Care. 2021 Dec. URL: http://www.g7.utoronto.ca/healthmins/G7-Open-Standards-and-Interoperability-Final-Report.pdf [accessed 2023-02-16]
43. Edwards A, Hollin I, Barry J, Kachnowski S. Barriers to cross--institutional health information exchange: a literature review. J Healthc Inf Manag 2010;24(3):22-34. [Medline: 20677469]
44. Ingebrigtsen T, Georgiou A, Clay-Williams R, Magrabi F, Hordern A, Prgomet M, et al. The impact of clinical leadership on health information technology adoption: systematic review. Int J Med Inform 2014 Jun;83(6):393-405. [doi: 10.1016/j.ijmedinf.2014.02.005] [Medline: 24656180]
45. Ross J, Stevenson F, Lau R, Murray E. Factors that influence the implementation of e-health: a systematic review of systematic reviews (an update). Implement Sci 2016 Oct 26;11(1):146 [FREE Full text] [doi: 10.1186/s13012-016-0510-7] [Medline: 27782832]
46. A plan for digital health and social care. Department of Health & Social Care. URL: https://www.gov.uk/government/publications/a-plan-for-digital-health-and-social-care/a-plan-for-digital-health-and-social-care [accessed 2023-02-16]
47. Elliott RA, Camacho E, Jankovic D, Sculpher MJ, Faria R. Economic analysis of the prevalence and clinical and economic burden of medication error in England. BMJ Qual Saf 2021 Feb;30(2):96-105. [doi: 10.1136/bmjqs-2019-010206] [Medline: 32527980]
48. Shirdel A, Pourreza A, Daemi A, Ahmadi B. Health-promoting services provided in pharmacies: a systematic review. J Educ Health Promot 2021;10:234 [FREE Full text] [doi: 10.4103/jehp.jehp_1374_20] [Medline: 34395671]
49. Community pharmacy contractual framework 2019-2024. NHS England. URL: https://www.england.nhs.uk/primary-care/pharmacy/community-pharmacy-contractual-framework/ [accessed 2023-02-16]
50. Hub and spoke dispensing. Department of Health & Social Care. 2022. URL: https://www.gov.uk/government/consultations/hub-and-spoke-dispensing/hub-and-spoke-dispensing [accessed 2022-08-03]

51. Pharmacy bodies highlight untenable funding and workforce challenges to Health Select Committee. Pharmaceutical Services Negotiating Committee. URL: https://psnc.org.uk/our-news/pharmacy-bodies-highlight-untenable-funding-and-workforce-challenges-to-health-select-committee/ [accessed 2022-08-03]
52. Alderwick H, Charlesworth A. A long term workforce plan for the English NHS. BMJ 2022 Apr 26;377:o1047. [doi: 10.1136/bmj.o1047] [Medline: 35474169]
53. Poon E, Trent Rosenbloom S, Zheng K. Health information technology and clinician burnout: current understanding, emerging solutions, and future directions. J Am Med Inform Assoc 2021 Apr 23;28(5):895-898 [FREE Full text] [doi: 10.1093/jamia/ocab058] [Medline: 33871016]
54. Bodenheimer T, Sinsky C. From triple to quadruple aim: care of the patient requires care of the provider. Ann Fam Med 2014;12(6):573-576 [FREE Full text] [doi: 10.1370/afm.1713] [Medline: 25384822]
55. Finlay L. Negotiating the swamp: the opportunity and challenge of reflexivity in research practice. Qual Res 2016 Aug 17;2(2):209-230. [doi: 10.1177/146879410200200205]

Abbreviations

dm+d: dictionary of medicines and devices
EHR: electronic health record
GNCR: Great North Care Record
GP: general practitioner
ICS: integrated care system
NENC: North East and North Cumbria
NHS: National Health Service
SNOMED: Systematized Nomenclature of Medicine

Edited by C Lovis; submitted 05.09.22; peer-reviewed by A Burgin, S Bacon, L Rotteau; comments to author 24.11.22; revised version received 05.01.23; accepted 25.01.23; published 10.03.23.

Please cite as:
Tolley C, Seymour H, Watson N, Nazar H, Heed J, Belshaw D
Barriers and Opportunities for the Use of Digital Tools in Medicines Optimization Across the Interfaces of Care: Stakeholder Interviews in the United Kingdom
JMIR Med Inform 2023;11:e42458
URL: https://medinform.jmir.org/2023/1/e42458
doi:10.2196/42458
PMID:36897631

Original Paper

Using the H2O Automatic Machine Learning Algorithms to Identify Predictors of Web-Based Medical Record Nonuse Among Patients in a Data-Rich Environment: Mixed Methods Study

Yang Chen[1], BSc; Xuejiao Liu[1], BEng; Lei Gao[2], LLM; Miao Zhu[3], PhD; Ben-Chang Shia[4,5], PhD; Mingchih Chen[4,5], PhD; Linglong Ye[6], PhD; Lei Qin[1,7], PhD

[1]School of Statistics, University of International Business and Economics, Beijing, China

[2]School of Law, University of International Business and Economics, Beijing, China

[3]School of Statistics, Huaqiao University, Xiamen, China

[4]Graduate Institute of Business Administration, College of Management, Fu Jen Catholic University, New Taipei City, Taiwan

[5]Artificial Intelligence Development Center, Fu Jen Catholic University, New Taipei City, Taiwan

[6]School of Public Affairs, Xiamen University, Xiamen, China

[7]Dong Fureng Institute of Economic and Social Development, Wuhan University, Wuhan, China

Corresponding Author:
Lei Qin, PhD
School of Statistics, University of International Business and Economics
No.10, Huixin Dongjie, Chaoyang District
Beijing, 100029
China
Phone: 86 01064491146
Email: qinlei@uibe.edu.cn

Abstract

Background: With the advent of electronic storage of medical records and the internet, patients can access web-based medical records. This has facilitated doctor-patient communication and built trust between them. However, many patients avoid using web-based medical records despite their greater availability and readability.

Objective: On the basis of demographic and individual behavioral characteristics, this study explores the predictors of web-based medical record nonuse among patients.

Methods: Data were collected from the National Cancer Institute 2019 to 2020 Health Information National Trends Survey. First, based on the data-rich environment, the chi-square test (categorical variables) and 2-tailed *t* tests (continuous variables) were performed on the response variables and the variables in the questionnaire. According to the test results, the variables were initially screened, and those that passed the test were selected for subsequent analysis. Second, participants were excluded from the study if any of the initially screened variables were missing. Third, the data obtained were modeled using 5 machine learning algorithms, namely, logistic regression, automatic generalized linear model, automatic random forest, automatic deep neural network, and automatic gradient boosting machine, to identify and investigate factors affecting web-based medical record nonuse. The aforementioned automatic machine learning algorithms were based on the R interface (R Foundation for Statistical Computing) of the *H2O* (H2O.ai) scalable machine learning platform. Finally, 5-fold cross-validation was adopted for 80% of the data set, which was used as the training data to determine hyperparameters of 5 algorithms, and 20% of the data set was used as the test data for model comparison.

Results: Among the 9072 respondents, 5409 (59.62%) had no experience using web-based medical records. Using the 5 algorithms, 29 variables were identified as crucial predictors of nonuse of web-based medical records. These 29 variables comprised 6 (21%) sociodemographic variables (age, BMI, race, marital status, education, and income) and 23 (79%) variables related to individual lifestyles and behavioral habits (such as electronic and internet use, individuals' health status and their level of health concern, etc). H2O's automatic machine learning methods have a high model accuracy. On the basis of the performance of the validation data set, the optimal model was the automatic random forest with the highest area under the curve in the validation set (88.52%) and the test set (82.87%).

Conclusions: When monitoring web-based medical record use trends, research should focus on social factors such as age, education, BMI, and marital status, as well as personal lifestyle and behavioral habits, including smoking, use of electronic devices and the internet, patients' personal health status, and their level of health concern. The use of electronic medical records can be targeted to specific patient groups, allowing more people to benefit from their usefulness.

(JMIR Med Inform 2023;11:e41576) doi:10.2196/41576

KEYWORDS

web-based medical record; predictors; H2O's automatic machine learning; Health Information National Trends Survey; HINTS; mobile phone

Introduction

Background

Regular review of self–medical records by patients can enhance patient-doctor communication and facilitate disease treatment. Effective communication can increase patient satisfaction, acceptance, adherence, and co-operation with the medical team. It can also improve a patient's physiological and functional status [1]. Conversely, poor communication between doctors and patients can lead to poor quality and continuity of care [2]. Therefore, ensuring good communication by recording, processing, and sharing health information with patients is a necessary and integral part of the health care process. Encouraging patients to use medical records can reduce unnecessary duplication of testing and treatment [3].

Before the advent of electronic medical records, traditional paper-based medical records written in technical language and comprising raw data were provided to health care professionals. However, such medical records can be worrying and confusing for patients. Consequently, clinical trials that provided written records to patients at the time reported that the use of medical records by patients had little success in enhancing communication and facilitating disease treatment [4-7]. However, with the advent of electronic storage of medical records and the internet, patients can be provided with web-based access to their medical records. Internet-accessible medical records may be particularly helpful to patients compared with centrally stored paper-based medical records. Patients can review web-based medical records repeatedly at their convenience. The readability optimization of web-based cases and the increasing popularity of internet medical information have made understanding web-based medical records easier for patients. Moreover, with the current COVID-19 pandemic, the use of web-based medical records may become more prevalent.

Studies have shown that providing patients with internet-accessible medical records may lead to modest benefits. For example, overall adherence to medical advice improved among patients using web-based medical records. A trend of improvement in satisfaction with doctor-patient communication has also been observed [8]. However, many patients avoid using web-based medical records despite their greater availability and readability. Historically, this was possibly because patients had little access to web-based medical records. For example, in 2013, only 3 in 10 patients gained access to medical records, and almost half of those who gained access viewed their web-based records at least once [9]. What factors influence the use of web-based medical records in the current population? This study explored the factors that influence people's nonuse of web-based medical records.

Some studies have applied traditional statistical methods to explore the relationship between certain factors and web-based medical record nonuse. For example, using univariable and multivariable regression models, Gerber et al [10] analyzed the use of MyChart (a personal health record portal for electronic medical record systems) among patients attending a National Cancer Institute–designated cancer center and predictors of MyChart use. Using data from the Health Information National Trends Survey (HINTS) cycle 3, Elkefi et al [9] applied descriptive statistics and chi-square tests to explore why patients tended to avoid using web-based medical records and compared patients' perceptions of web-based medical records based on demographics and cancer diagnoses. On the basis of 2017 to 2018 HINTS data, Patel and Johnson [11] used descriptive statistics and hypothesis testing to assess individuals' access, viewing, and use of their web-based medical records and the use of smartphone health apps and other electronic devices in 2017 and 2018. Trivedi et al [12] used multivariable logistic regression (LR) analyses to examine the association between sociodemographic and health care–related factors on being offered access to web-based medical records and accessing web-based medical records and cited reasons for not accessing web-based medical records. These studies used traditional and relatively simple statistical methods, and the selection of predictors has certain limitations. As in the research by Elkefi et al [9], predictors relate only to demographic variables and cancer diagnoses. Screening of predictors based on a data-rich environment can be optimized.

With the rapid development of artificial intelligence, machine learning methods have received increasing attention. Machine learning algorithms are used in a wide variety of applications, such as in medicine and health care, where it is difficult or unfeasible to develop conventional algorithms for necessary tasks [13]. Compared with traditional regression-based statistical methods, machine learning is data-driven and has the advantage of not assuming the distribution and relationship of predictors, and machine learning algorithms are good at handling data that are multidimensional and multivariety. Deep learning is a step forward, which makes feature engineering part of the learning task, reducing the algorithm's dependence on feature engineering. However, the parameters of the machine learning method greatly influence model accuracy. Incorrect parameter selection and a small sample size can both lead to reduced model performance. Some parameters (such as the number of trees, learning rate, and number of leaf nodes in the random forest

method) determine the structure and training method of the model, which affects prediction performance. To take full advantage of the relevant machine learning algorithms, an appropriate strategy must be developed to determine the parameters.

Objectives

In this study, explanatory variables were chosen based on a data-rich environment. Data for this study were collected from the National Cancer Institute 2019 to 2020 HINTS. The HINTS regularly collects nationally representative data about the American public's knowledge of, attitudes toward, and use of cancer- and health-related information; therefore, this study is based on the relevant background in the United States. We used almost all the questions in the questionnaire as possible predictors, thus avoiding the subjectivity of manual screening. To resolve the parameter selection problem of machine learning algorithms, this study adopted the current popular *H2O* (H2O.ai) automatic machine learning algorithms to realize the automation of the entire process, from construction to the application of the machine learning model. At the same time, we also used the traditional statistical method of LR. To the best of our knowledge, this is the first study that applied a range of H2O's automatic machine learning algorithms to such a large representative sample based on a data-rich environment. We implemented a combination of H2O's automatic machine learning methods and a data-rich environment. Predictors of web-based medical record nonuse were identified based on the results of the H2O automatic machine learning methods.

Methods

Data Source

Data for this study were collected from the National Cancer Institute 2019 to 2020 HINTS. The HINTS regularly collects nationally representative data about the American public's knowledge of, attitudes toward, and use of cancer- and health-related information. Survey researchers use the data to understand how adults (aged ≥18 years) use different communication channels, including the internet, to obtain vital health information for themselves and their loved ones. This study analyzed merged data from cycles 3 to 4. Data from cycle 3 were collected between January 2019 and May 2019, and those from cycle 4 were collected from February 2019 to June 2019. We screened the respondents based on the target-dependent variable (ie, web-based medical record nonuse), leaving respondents with no missing values in the target-dependent variable. Finally, 9072 respondents were screened.

Ethical Considerations

The HINTS administration was approved by the institutional review board at Westat Inc and deemed exempt by the National Institutes of Health Office of Human Subjects Research. This exemption also extends to this study. HINTS data are available for public use. Additional information on the survey design is available on the HINTS website.

Statistical Analysis

Explanatory variables were selected in this study based on a data-rich environment, and all questions in the questionnaire that could be answered by all participants were selected (P_0=141; variables that could only be answered by a specific group were not considered, such as questions only for females, eg, whether they had been screened for cervical cancer). The sociodemographic characteristics and other relevant variables of individuals who had or had not used web-based medical records were compared using chi-square tests for categorical variables and 2-tailed *t* tests for continuous variables. According to the results of the aforementioned statistical tests, significant variables were selected. Preliminary screening of variables was completed (P_1=49; some variables were merged and answers were regrouped; refer to Multimedia Appendix 1 for details). Samples with missing values for the preliminary screened variables were excluded, and accordingly, a total of 4827 samples were obtained. On the basis of these samples, 5 algorithms were used for modeling: LR, automatic generalized linear model (auto-GLM), automatic random forest, automatic deep neural network (auto–deep learning), and automatic gradient boosting machine (auto-GBM). Of them, LR is a traditional statistical method, and the last 4 are automated machine learning algorithms based on the R interface (R Foundation for Statistical Computing) of the H2O extensible machine learning platform.

We divided the data set as follows: 80% of the data were used as the training set, and 20% of the data were used as the test set. We used the method of 5-fold cross-validation on the training data to determine hyperparameters and used the selected optimal hyperparameters to fit the model using all training data and make predictions on the test set. To evaluate the predictive accuracy of the models, we reported the accuracy, precision, recall, F_1-score, and area under the curve (AUC) of the validation set (validation set results for 5-fold cross-validation in the training set) and test set. The LR model selected predictors through backward selection and stepwise regression. The relative effects of the predictors in the LR model were measured based on crude odds ratios (ORs), whereas the variability and significance were assessed based on CIs and the corresponding P values. Variable importance values were used in the other 4 H2O automatic machine learning classification algorithms to identify predictors (variables with higher importance indexes were screened as predictors, and a 5-fold cross-validation method was used to select important variables by using all data). All statistical analyses were performed using the R software (version 4.1.2). In this study, P<.05 was considered statistically significant.

Measures

Nonuse of Web-Based Medical Records

Web-based medical records are used to organize processes in clinical and outpatient settings and forge doctor-patient communication that establishes mutual understanding and trust. The variable "nonuse of web-based medical records" in this study was calculated based on the following question in the HINTS: "How many times did you access your web-based

medical record in the last 12 months?" We used this question to identify users and nonusers of web-based medical records. The respondents who reported accessing their web-based medical records at least once were coded as users, and those who reported accessing their records 0 times were coded as nonusers.

Demographic and Other Related Variables

Demographic variables of interest (dichotomized for analysis) included sex (male and female), race and ethnicity (non-Hispanic White and racial and ethnic minority group), education (high school or lower and more than high school), income ranges (<US $20,000 and ≥US $20,000), area (nonmetropolitan and metropolitan), and marital status (married and not married), as well as numerical demographic variables, including age (continuous years) and BMI.

For further analysis, we selected as many variables as possible from the HINTS database to identify their relationship with the use of web-based medical records. Statistical tests were performed on almost all variables in the questionnaire, including chi-square tests for categorical variables and 2-tailed *t* tests for continuous variables. The variables that passed the significance test were used as potential predictors, as follows (consistent with the question blocks in the questionnaire): 6 variables, such as *Confidence in access to health information*, in part A (looking for health information); 6 variables, such as *Internet use*, in part B (using the internet to find information); 2 variables, such as *Have regular health providers*, in part C (your health care); 3 variables, such as *Health provider maintain MR (medical record)*, in part D (medical records); *Care for someone* in part E (caregiving); 7 variables, such as *General health*, in part F (your overall health); 2 variables, such as *Notice calorie information*, in part G (health and nutrition); 2 variables, such as *Exercise days per week*, in part H (physical activity and exercise); 3 variables, such as *Smoke*, in part K (tobacco products; this part of the questionnaire was about the respondents' consumption of tobacco products); 3 variables, such as *Ever tested colon cancer*, in part L (cancer screening and awareness); *Ever had cancer* in part M (your cancer history); 4 variables, such as *Everything cause cancer*, in part N (beliefs about cancer); and 1 numerical variable, *Sitting time per day*.

Specific variables and their descriptive statistics are shown in Multimedia Appendix 2, and Multimedia Appendix 1 lists details of some of the aforementioned variables, including demographic variables and variables adjusted for research needs with the readjustment information.

Machine Learning Methods

LR Model

LR is a generalized linear regression analysis model that is part of supervised learning in machine learning. LR usually uses numerical or categorical independent variables x_1, x_2,...,x_n to predict the value of the categorical dependent variable y to determine the probability that y belongs to a particular category [14].

$$p(y = 1 \mid x_1, x_2,...,x_n) = (e^{\beta_0 + \beta_1 x_1 + ... + \beta_n x_n})/(1 + e^{\beta_0 + \beta_1 x_1 + ... + \beta_n x_n}) \quad \textbf{(1)}$$

The OR expresses the ratio between the probability p that the dependent variable y is 1 and the probability $1 - p$ that the dependent variable y is 0. The OR is related to the interpretability of LR. When x_i is increased by 1, the odds become the original $e^{\beta i}$ times.

$$logit(p) = \ln(p/1-p) = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n \quad \textbf{(2)}$$

In the aforementioned formula, β_1, β_2,..., β_n are the coefficients that measure the contribution of the independent variables x_1, x_2,...,x_n to y. If the coefficient β is positive, $e^{\beta}>1$ and the factor have a direct correlation with y, whereas if β is negative, e^{β} is between 0 and 1.

H2O's Auto-GLM

Generalized linear models (GLMs) were proposed and published by Nelder and Wedderburn [15] in 1972. It is a modeling method that can solve the problem that ordinary linear regression models cannot handle discrete dependent variables. GLM is an extension of the linear model and establishes the relationship between the mathematical expectation of the response variable and the linear combination of predictor variables through a link function. In this study, 5-fold cross-validation was adopted on the data set to select the hyperparameters of the model, and the selection range of the regularization parameter was (0, 1). Ridge regression (α=0) was used in the regression of the GLM for the variable selection and final classification. The importance of the variable was judged according to the "absolute value of the normalization coefficient" indicator; the larger the value, the greater the importance of the variable. This study used the "h2o.glm" function in the "h2o" package to build a GLM for the classification of web-based medical record use.

H2O is an open-source, in-memory, distributed, fast, and scalable machine learning and predictive analytics platform that allows users to build machine learning models on data and avoid the tedious process of manual hyperparameter tuning. H2O supports traditional (or "Cartesian") grid searches. In a Cartesian grid search, users specify a set of values for each hyperparameter that they want to search, and H2O trains a model for every combination of the hyperparameter values. This means that, if we have 3 hyperparameters and specify 5, 10, and 2 values for each, the grid will contain a total of 5 × 10 × 2 = 100 models. After the grid search is complete, the user can query the grid object and sort the models by a specific performance metric (eg, "AUC") and select the locally best model within the specified parameter range.

H2O's Automatic Random Forest

The random forest is a multivariate statistical technique that considers an ensemble (forest) of trees for efficiency and predictive power [16]. Random forest uses a bagging technique (bootstrap aggregation) to select resamples randomly and choose a random sample of variables at each tree node as the training data set for model calibration. As the random selection of the training data set may affect the model's results, a large set of trees is applied to guarantee model stability. In this study, the

selection range of the number of trees was between 100 and 500, the selection range of the number of variables in the variable selection set at the node of the tree was approximately $p^{0.5}$ (p is the total number of variables), and the maximum tree depth was selected from 10 to 30. When selecting variables, the parameters of the final model selected under 5-fold cross-validation were as follows: the number of trees was 150, the number of variables contained in the variable selection set at the node of the tree was 7, and the maximum depth of the tree was 10. The importance of the variable was judged according to the "mean decrease gini" indicator, where the larger the value, the greater the importance of the variable. When fitting the model, the final parameters were as follows: the number of trees was 300, the number of variables included in the variable selection set at the node of the tree was 2, and the maximum depth of the tree was 10. Model-fitting processes were implemented using the "h2o.randomForest" function in the "h2o" package. Parameter tuning was implemented using the "h2o.grid" function in the "h2o" package by grid searching for parameters.

H2O's Auto–Deep Learning

The concept of deep learning originates from the study of artificial neural networks, and a multilayer perceptron with multiple hidden layers is a basic deep learning structure. Deep learning algorithms try to identify potential relationships in a data set by mimicking human brain functions. Similar to the human brain structure, deep learning models consist of neurons in complex and nonlinear forms. Deep learning models have 3 basic types of layers: input, hidden, and output layers. Each neuron in the current layer is connected to the input signal of each neuron in the previous layer. In each connection process, the signal from the previous layer is multiplied by a weight, and a bias is added and then passed through a nonlinear activation function through multiple composites of simple nonlinear functions to achieve a complex input space–to–output space map. In this study, the input values were observations of 49 variables, and the output value was the probability of the use of web-based medical records. When training the model, the number of hidden layers was 2 to 3; the number of nodes in the first layer was between 100 and 200; the number of nodes in the second layer was between 50 and 100; and the number of nodes in the third layer was 5, if any. The activation function was selected from the rectifier and rectifier with dropout; dropout ratio defaults to 0.5. The deep learning model chosen using 5-fold cross-validation to be applied for selecting variables contained 3 hidden layers, each with 100, 50, and 5 nodes. When training the model, a 50% random dropout of the nodes was set to prevent overfitting. The variable importance of the model was measured using the combination of absolute values of the coefficients. The final model for classification contained 3 hidden layers, each with 200, 50, and 5 nodes with 50% random dropout, which provided the highest mean AUC in the test set of the model in 5-fold cross-validation. In this study, we used the "h2o.deeplearning" function in the "h2o" package to realize the deep learning algorithm.

H2O's Auto-GBM

The gradient boosting machine (GBM) algorithm is a type of boosting algorithm. GBM is a model that trains decision trees sequentially. Each decision tree is based on the errors of the previous tree. The core idea is to generate various weak learners in series, and the goal of each weak learner is to fit the negative gradient of the loss function of the previous accumulated model. After adding the weak learner, it enables the accumulated model loss to decrease along the negative gradient direction. It uses different weights to linearly combine the basic learners to ensure that learners with greater performance can obtain larger weights. The most commonly used base learners are tree models. Variable importance is determined by calculating the relative influence of each variable: whether that variable is selected to split during the tree-building process and how much the squared error improves or decreases as a result. We used 5-fold cross-validation in both the variable selection and classification model-fitting procedures, which would provide indicators for the selection of optimal hyperparameters. In the process of using a grid search for hyperparameter optimization, it is necessary to train the models under different hyperparameter specifications and evaluate the goodness of fit of the models under these specifications through 5-fold cross-validation. The number of trees ranged from 100 to 500, and the final parameter was set to 300 in both the variable selection and classification model-fitting procedures, which promises a balance in the training and test set errors in 5-fold cross-validation. In addition, the selection of the learning rate ranged from 0.01 to 0.10, and the final parameter was set to 0.01. The maximum depth was between 10 and 30, and the final parameter was set to 10 in both the variable selection and classification model-fitting procedures, which implements a trade-off between model bias and model variance. This study used the "h2o.gbm" function in the "h2o" package to build a GBM for the use of web-based medical record classification problems. Parameter tuning was implemented using the "h2o.grid" function in the "h2o" package by grid searching for parameters.

Results

Descriptive Statistics

The merged data sets from HINTS cycles 3 and 4 yielded a sample of 9072 respondents, including 5409 (59.62%) nonusers and 3663 (40.38%) users of web-based medical records. Multimedia Appendix 2 presents the frequencies and proportions of the variables. The chi-square test of categorical variables and the 2-tailed t test of continuous variables showed significant differences in some variables between nonusers and users of web-based medical records (P<.05). Among the categorical variables, respondents who chose the following options comprised a significantly higher proportion (P<.05) of the group not using web-based medical records: "male," "trust information about health or medical topics from religious organizations and leaders," "have no drink," and "smoke more." For example, in this group, male individuals accounted for 45.23% (2196/4855) of the respondents, whereas in the group using web-based medical records, the percentage decreased to 38.65% (1331/3444). The same was true for the other aforementioned variables: "trust information about health or medical topics from

religious organizations and leaders" (1501/4934, 30.42% vs 824/3573, 23.06%), "have no drink" (2623/4692, 55.9% vs 1575/3421, 46.04%), and "smoke more" (824/3573, 39.38% vs 1264/3630, 34.82%). However, those choosing "Non-Hispanic White" (3492/4885, 71.48% vs 2686/3488, 77.01%), "have higher education level" (3471/5212, 66.6% vs 3160/3598, 87.83%), "have higher income level" (3569/4741, 75.28% vs 3042/3349, 90.83%), "in marriage" (2481/5198, 47.73% vs 2252/3600, 62.56%), "ever looked for information about cancer" (2378/5344, 44.50% vs 2429/3646, 66.62%), "use Internet" (3841/5379, 71.41% vs 3501/3650, 95.92%), "use electronic" (3752/5346, 70.18% vs 3572/3642, 98.08%), "use Internet for health use" (3337/5289, 63.09% vs 3111/3636, 85.56%), "have regular healthcare provider" (3285/5293, 62.06% vs 2930/3625, 80.83%), "caring someone" (723/5216, 13.86% vs 656 /3604, 18.20%), "general health relative good" (4343/5331, 81.47% vs 3169/3625, 87.42%), "high confidence in ability to take good care of own health" (5036/5337, 94.36% vs 3503/3632, 96.45%), and others were significantly higher (P<.05) in the group using web-based medical records. Among the numeric variables, mean values of age were significantly higher (P<.05) in the group not using web-based medical records, whereas time spent sitting was significantly higher (P<.05) in the group using web-based medical records. The 2-tailed t test of continuous variables also showed no significant difference (P>.05) in some variables between individuals who had and had not used web-based medical records. In other words, the proportions of these variables were similar between the 2 groups. As for BMI, the average value in both groups was approximately 28.5 (SD 0.1).

Machine Learning Model Results

As shown in Tables 1 and Table 2, a total of 29 predictors of nonuse of web-based medical records variables (*Age, Sitting time per day, BMI, Confidence in access to health information, education, Electronic means use, Ever tested colon cancer, Everything cause cancer, Number of visits to health provider, Have electronic device, income, Obesity affects cancer onset, Social media use, Little interest, Marital status, Offered access to MR by health provider, Offered access to MR by health insurer, Health provider maintain MR, race, Have regular health providers, Seek cancer information, Shared health information, Smoke, Exercise days per week, Strength training days per week, Trust doctor, Trust religious organizations, Internet use, and Electronic wearable device use*) were selected in all the 5 algorithms, and 7 variables (*Age, Electronic means use, Number of visits to health provider, Offered access to MR by health provider, Offered access to MR by health insurer, Health provider maintain MR, and Internet use*) were selected simultaneously in the 5 algorithms.

Table 1. Predictors of nonuse of web-based medical records (MRs) using the logistic regression algorithm.

Predictor	OR[a] (95% CI)
Race (reference: non-Hispanic White)	1.04 (0.90-1.19)
Education (reference: high school or lower)	0.32 (0.27-0.37)
Income (reference: <US $20,000)	0.35 (0.29-0.43)
Marital status (reference: not married)	0.60 (0.54-0.68)
Trust doctor[b] (reference: low_level)	0.73 (0.53-0.99)
Trust religious organization[b] (reference: low_level)	1.31 (1.15-1.50)
Internet use (reference: no)	0.12 (0.09-0.16)
Electronic means use (reference: no)	0.05 (0.03-0.07)
Electronic wearable device use (reference: no)	0.40 (0.35-0.45)
Shared health information (reference: N/A[c])	
No	0.71 (0.58-0.87)
Yes	0.26 (0.20-0.33)
Social media use (reference: no)	0.37 (0.32-0.43)
Have regular health providers[d] (reference: no)	0.36 (0.32-0.42)
Number of visits to health provider[e] (reference: none)	
1 time	0.24 (0.18-0.33)
2 times	0.19 (0.15-0.26)
3 times	0.15 (0.12-0.21)
4 times	0.14 (0.11-0.19)
5-9 times	0.10 (0.08-0.14)
≥10 times	0.10 (0.08-0.14)
Health provider maintain MR (reference: no)	
Yes	0.11 (0.06-0.22)
Don’t know	1.14 (0.56-2.30)
Offered access to MR by health provider[f] (reference: no)	
Yes	0.03 (0.03-0.04)
Don’t know	0.83 (0.57-1.22)
Offered access to MR by health insurer[f] (reference: no)	
Yes	0.21 (0.19-0.25)
Don’t know	0.82 (0.71-0.95)
Strength training days per week (reference: none)	
1-3 days per week	0.63 (0.56-0.71)
4-7 days per week	0.84 (0.70-1.01)
Ever tested colon cancer (reference: no)	0.74 (0.66-0.83)
Age	1.01 (1.00-1.01)
BMI	0.99 (0.99-1.00)

[a]OR: odds ratio.

[b]In general, how much would you trust information about cancer from a doctor/government health agencies/charitable organizations/religious organizations and leaders? (Supplement to the variable-related questions in the survey).

[c]N/A: not applicable.

[d]Not including psychiatrists and other mental health professionals, is there a particular doctor, nurse, or other health professional that you see most often? (Supplement to the variable-related questions in the survey).

[e]In the past 12 months, not counting times you went to an emergency room, how many times did you go to a doctor, nurse, or other health professional to get care for yourself? (Supplement to the variable-related questions in the survey).

[f]Have you ever been offered online access to your medical records by your health care provider/health insurer? (Supplement to the variable-related questions in the survey).

Table 1 shows significant predictors of nonuse of web-based medical records in LR (P<.05). The variables in LR were screened using 2 methods: backward selection and stepwise regression. The results obtained using the 2 variable selection methods were consistent, and 20 significant variables were finally selected.

The results of LR showed that sociodemographic indicators, such as age, BMI, education, marital status, income, and race, significantly affected the nonuse of web-based medical records, whereas sex and area had no significant effect on the prediction of nonuse of web-based medical records. On the basis of sociodemographic variables, people who were relatively older (OR 1.01, 95% CI 1.00-1.01), had a relatively lower BMI (OR 0.99, 95% CI 0.99-1.00), had relatively lower education (higher education OR 0.32, 95% CI 0.27-0.37), were not married (married OR 0.60, 95% CI 0.54-0.68), had a lower income (higher income OR 0.35, 95% CI 0.29-0.43), and belonged to racial and ethnic minority groups (OR 1.04, 95% CI 0.9-1.19) were more likely to not use web-based medical records. People who did not often access the internet or send and receive emails (OR 0.12, 95% CI 0.09-0.16), had not used a computer or smartphone to inquire about medical information in the past 12 months (OR 0.05, 95% CI 0.03-0.07), did not often use a wearable device to track health (OR 0.40, 95% CI 0.35-0.45), did not share health information from an electronic monitoring device or smartphone with a health professional in the previous 12 months, and people who had not used social media in the last 12 months (OR 0.37, 95% CI 0.32-0.43) were more inclined not to use web-based medical records. Moreover, those who were less concerned about their own health were more likely to not use web-based medical records. People who were more likely to not use web-based medical records tended to be those who did not see a particular doctor or health care professional frequently (OR 0.36, 95% CI 0.32-0.42); had not gone to a doctor, nurse, or other health professional to receive care in the last 12 months; did not have doctors or other health care providers maintain their medical records in a computerized system (OR 0.11, 95% CI 0.06-0.22); were not offered web-based access to their medical records by the health care provider (OR 0.03, 95% CI 0.03-0.04); were not offered web-based access to their medical records by the health insurer (OR 0.21, 95% CI 0.19-0.25); and did not perform leisure-time physical activities specifically designed to strengthen muscles. People who strongly trusted information about health or medical topics from religious organizations and leaders (OR 1.31, 95% CI 1.15-1.50), trusted information about health or medical topics from a doctor only a little (OR 0.73, 95% CI 0.53-0.99), and did not check for colon cancer (OR 0.74, 95% CI 0.66-0.83) were more inclined to not use web-based medical records.

Table 2. Predictors of web-based medical record (MR) nonuse built using automatic generalized linear model (auto-GLM), automatic random forest, auto–deep learning, and automatic gradient boosting machine (auto-GBM).

Model and predictor	Importance scores
Auto-GLM	
Offered access to MR by health provider[a]	100
Electronic means use	47.41
Health provider maintain MR	22.72
Number of visits to health provider[b]	21.63
Age	17.11
Offered access to MR by health insurer[a]	14.94
Electronic wearable device use	13.91
Have regular health providers[c]	13.10
Shared health information	12.06
Internet use	10.97
Ever tested colon cancer	9.64
Social media use	9.30
Income	6.45
Education	6.27
Race	6.23
Automatic random forest	
Offered access to MR by health provider	100
Electronic means use	19.15
Offered access to MR by health insurer	19.02
Health provider maintain MR	16.03
Number of visits to health provider	10.20
Internet use	6.88
Have regular health providers	5.97
Age	5.43
Shared health information	4.71
Electronic wearable device use	4.58
Sitting time per day	4.29
BMI	4.15
Have electronic device	4.13
Education	2.96
Sitting time per day	2.83
Auto–deep learning	
Offered access to MR by health provider	100
Electronic means use	51.18
Health provider maintain MR	40.19
Internet use	34.94
Offered access to MR by health insurer	34.41
Number of visits to health provider	33.24
Little interest[d]	33.02

Model and predictor	Importance scores
Social media use	32.59
Smoke	32.22
Confidence in access to health information	31.86
Race	31.80
Sitting time per day	31.71
Age	31.62
Obesity affects cancer onset	30.91
Have electronic device	30.78
Auto-GBM	
Offered access to MR by health provider	100
Electronic means use	15.87
Number of visits to health provider	6.84
Age	3.97
Offered access to MR by health insurer	3.95
Sitting time per day	3.08
Electronic wearable device use	2.48
Shared health information	2.08
Internet use	1.80
BMI	1.49
Health provider maintain medical records	1.46
Have electronic device	1.28
Have regular health providers	1.23
Exercise days per week	1.13
Everything cause cancer	0.97

[a]Have you ever been offered online access to your medical records by your health care provider/health insurer? (Supplement to the variable-related questions in the survey).

[b]In the past 12 months, not counting times you went to an emergency room, how many times did you go to a doctor, nurse, or other health professional to get care for yourself? (Supplement to the variable-related questions in the survey).

[c]Not including psychiatrists and other mental health professionals, is there a particular doctor, nurse, or other health professional that you see most often? (Supplement to the variable-related questions in the survey).

[d]Over the past 2 weeks, how often have you been bothered by little interest or pleasure in doing things? (Supplement to the variable-related questions in the survey).

The essence of this study is a binary classification problem of judging whether individuals have used web-based medical records based on a set of inputs, such as education and income. Therefore, we evaluated the 5 methods using a series of evaluation metrics commonly used for classification algorithms. Table 3 presents the accuracy, precision, recall, F_1-score, and AUC values for the 5 machine learning methods on the validation and test sets. Accuracy is a metric of a classification model that measures the percentage of correct classification accounts for the total number of classifications. Precision is the proportion of correctly predicted positives to all predicted positives, whereas recall is the proportion of correctly predicted positives to all actual positives. The F_1-score is the harmonic mean of precision and recall. From the results of the verification set, LR had 3 indicators that performed best, with an accuracy of 83.35%, a recall of 88.97%, and an F_1-score of 83.59%. However, the performance of LR on the test set was inferior to that of machine learning methods. In the test set, auto-GLM had the highest accuracy (82.49%), auto-GBM had the highest precision (79.73%), and automatic random forest had the highest recall (87.15%) and AUC (82.87%). AUC is not affected by the classification threshold and data distribution and, thus, reflects the overall classification power of the model. Automatic random forest had the highest AUC in both the validation (88.52%) and test (82.87%) sets. Therefore, in general, we were more inclined to choose the automatic random forest as the optimal model for predicting web-based medical record nonuse. Equations 3 to 6 provide the evaluation formulas for the machine learning models.

$$Accuracy = (TP\ [True\ Positive] + TN\ [True\ Negative])/(TP + TN + FP\ [False\ Positive] + FN\ [False\ Negative]) \quad (3)$$

$$Precision = TP/(TP + FP) \quad (4)$$

$$Recall = TP/(TP + FN) \quad (5)$$

$$F1\text{-}score = 2 \times (Recall \times Precision)/(Recall + Precision) \quad (6)$$

Table 3. Correct classification metrics for each machine learning method.

Criterion	LR[a]		Auto-GLM[b]		Automatic random forest		Auto–deep learning		Auto-GBM[c]	
	Validation	Test	Validation	Test	Validation	Test	Validation	Test	Validation	Test
Accuracy, %	83.35	81.47	82.17	82.49	82.48	82.38	82.57	80.93	81.32	79.69
Precision, %	78.86	76.26	85.09	77.73	86.1	76.43	89.52	74.3	85.38	79.73
Recall, %	88.97	86.7	79.76	84.81	79.04	87.15	75.01	87.15	77.78	76.96
F_1-score, %	83.59	82.01	82.34	81.11	82.3	81.44	81.57	80.22	81.31	78.32
AUC[d], %	82.29	81.8	88.46	82.72	88.52	82.87	87.54	81.56	87.68	79.57

[a]LR: logistic regression.

[b]Auto-GLM: automatic generalized linear model.

[c]Auto-GBM: automatic gradient boosting machine.

[d]AUC: area under the curve.

Discussion

Principal Findings

Effective communication is key for delivering high-quality health care services [1]. Ensuring good communication by recording, processing, and sharing health information with patients is integral to the health care process. At present, the development of information and communications technology has allowed for the realization of web-based medical records, but because of some situations, web-based medical records are still not fully popular among patients. On the basis of demographic and individual behavioral characteristics, this study explored the predictors of web-based medical record nonuse.

We conducted a comprehensive assessment of the effects of individual sociodemographic characteristics, lifestyle, and behavioral habits on nonuse of web-based medical records. Although generalizing the factors that influence individuals' nonuse of web-based medical records was difficult, based on the survey data, this study conducted in a data-rich variable selection environment demonstrated that an individual's nonuse of web-based medical records is related to their sociodemographic characteristics, such as age, lifestyle, behavioral habits, and attention to health problems.

To date, numerous studies on the use of web-based medical records by patients have been based on surveys wherein data were gathered using a semistructured interview approach [17]. Using data from the National Cancer Institute 2019 to 2020 HINTS database, we applied the following question—"How many times did you access your web-based medical record in the last 12 months?"—to determine whether a person used web-based medical records. Our analysis showed that 59.62% (5409/9072) of the population in the survey samples had no experience using web-based medical records.

This study used 5 algorithms—LR, auto-GLM, automatic random forest, auto–deep learning, and auto-GBM—to identify and investigate factors affecting individuals' nonuse of web-based medical records. Of them, LR is a traditional statistical method, and the latter 4 algorithms are part of H2O's automatic parameterization methods. A total of 29 influencing variables concerning the use of web-based medical records were selected based on coefficient significance in the LR model and the variable importance indicators in the other 4 methods. Many well-established determinants were also identified as proof of concept for our analytical approach, such as sociodemographic characteristics [9]. Although nonlinear and ensemble algorithms exhibit better predictive performance than traditional parametric models, they are less interpretable [18]. Therefore, predictors determined by such algorithms should be evaluated in conjunction with relevant research evidence.

This study showed that sociodemographic indicators, such as age, BMI, race, marital status, education, and income, significantly affected the nonuse of web-based medical records, whereas sex and area did not significantly affect the prediction of the nonuse of web-based medical records. In addition, predictors involving personal lifestyle and behavioral habits, such as smoking, electronic device use, and internet use, also played essential roles in predicting the nonuse of web-based medical records. Finally, individuals' health status and their level of health concern were also associated with the nonuse of web-based medical records. Of course, some regions or units do not provide web-based access to medical records, or these are not maintained by health care providers, which may directly limit the use of web-based medical records.

On the basis of some of the conclusions of this study, recommendations can be made to promote the widespread use of web-based medical records. The use of electronic equipment is also a factor affecting the use of web-based medical records. People who are not accustomed to using electronic devices, such as mobile phones and computers, generally do not access their web-based medical records. The reason for avoiding web-based medical records may not be the disadvantage of web-based medical records itself but the resistance to electronic products, which is more common in older adults. However, older adults are more likely to become sick, so web-based medical records for this group are also a direction that needs special attention and development, such as building an interface

that is friendly to the older adult population and keeping the internet-accessible interface simple and clear. Web-based medical records are often used by people with more related health problems. A study by the Office of the National Coordinator of Health Information Technology found that individuals may not realize the value of accessing their web-based medical records until they have a medical need. Given that the patient record request process can be time consuming, it may be more beneficial to have access to a person's data in advance of an urgent health need. Therefore, popularizing health knowledge to the public and increasing the public's attention to health information can increase the public's demand for health-related information to a certain extent, thereby promoting the use of web-based medical records.

This research can provide a theoretical basis for predicting individual web-based medical record use. On the basis of the predictors of people not using web-based medical records selected by machine learning algorithms, individuals who do not use such records can be identified in advance, and use of web-based medical records can be promoted among them. Thus, this would provide more effective doctor-patient communication and better health care services.

Limitations

This study has some limitations. First, this study explored the influencing factors of nonuse of web-based medical records and discussed the correlation between each influencing factor and the target variable but did not involve the study of the influence path and influence mechanism. We considered conducting a causal analysis, but the cross-sectional survey design of the HINTS prevented us from making or testing causal claims. Second, the relationships between the selected variables were not studied further, and the screened important predictors may have certain collinearity. Third, the data were obtained using a self-report questionnaire. Therefore, we did not obtain detailed information on the nonuse of web-based medical records, and self-report bias may have affected the results. Finally, patient access to web-based medical records varies from country to country, and cultural background also has a strong impact on medical services. This study used HINTS data from the United States, and the conclusions may not be generalizable to other countries.

Comparison With Prior Work

The use of web-based medical records can enhance patient participation and co-operation in disease treatment and enhance doctor-patient communication to promote disease treatment. This is evidenced in the study by Stewart et al [19], whose research took patients with diabetes as the research object and found that patient portals support engagement by facilitating patient access to their health information and facilitating patient-provider communication. With the advancement of internet technology and the popularization of electronic products, the use of web-based medical records has become more convenient, but its penetration rate is still not high. Therefore, it has become a research hot spot to explore the characteristics and differences between users and nonusers of web-based medical records and identify the influencing factors of low use rate.

There are many studies based on HINTS data, such as that by Anthony et al [20], who used data from the 2017 HINTS to estimate 2 separate multivariable LR models to predict the factors associated with not having been offered access and those associated with not using a portal. On the basis of the 2017 to 2018 HINTS data, Patel and Johnson [11] used descriptive statistics and hypothesis testing to assess individuals' access, viewing, and use of their web-based medical records and the use of smartphone health apps and other electronic devices. Trivedi et al [12] used multivariable LR analyses to examine the association between sociodemographic and health care–related factors on being offered access to web-based medical records and accessing web-based medical records and cited reasons for not accessing web-based medical records. Hong et al [21] used LR to investigate the trend of patient portal use in the general population and the barriers to adoption. The aforementioned studies are all based on traditional statistical methods such as LR [12,20,21] and hypothesis testing related to association analysis [11], and the selection of influencing factors and the conclusions drawn only involve demographic variables and some variables of interest, whereas our research combines traditional statistical methods and machine learning methods to find predictors of web-based medical record use based on variable-rich environments. Furthermore, the results of machine learning methods also provide variable importance scores and rankings, which have also not been covered in previous studies. To the best of our knowledge, this is the first study to apply a range of H2O's automatic machine learning algorithms to a nationally representative sample for optimizing the classification of web-based medical record nonuse. Compared with previous studies, we found that personal lifestyle and behavioral habits as well as individuals' health status and their level of health concern significantly affect web-based medical record nonuse.

Conclusions

Using data from the National Cancer Institute 2019 to 2020 HINTS database, this study applied 5 machine learning algorithms—LR (linear), auto-GLM, automatic random forest, auto–deep learning, and auto-GBM—to identify and investigate the factors that affect whether individuals use web-based medical records. Using these 5 models, 29 variables were identified as crucial predictors of nonuse of web-based medical records. When monitoring web-based medical record use trends, research should consider social factors such as age, education, BMI, and marital status, as well as personal lifestyle and behavioral habits, including smoking, use of electronic devices and the internet, patients' personal health status, and their level of health concern. The use of electronic medical records can be targeted to specific patient groups, allowing more people to benefit from their usefulness.

The main contributions of this study are as follows: (1) using authoritative data, potential predictors were selected based on a data-rich environment, involving more comprehensive variables and avoiding unnecessary subjectivity, and (2) the key parameters of the machine learning methods considerably influenced the accuracy of the model. In this study, H2O's automatic parameter selection methods were introduced to optimize the key parameters of the model. Compared with

traditional machine learning algorithms, the H2O automatic machine learning methods effectively improved model performance.

Acknowledgments

This research was funded by the Youth Project of National Social Science Fund of China (grant 21CTJ008).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Variables extracted from the Health Information National Trends Survey database for research.
[DOCX File , 20 KB - medinform_v11i1e41576_app1.docx]

Multimedia Appendix 2
Distribution of characteristics of variables in the Health Information National Trends Survey database (N=9072).
[DOC File , 263 KB - medinform_v11i1e41576_app2.doc]

References

1. Norouzinia R, Aghabarari M, Shiri M, Karimi M, Samami E. Communication barriers perceived by nurses and patients. Glob J Health Sci 2015 Sep 28;8(6):65-74 [FREE Full text] [doi: 10.5539/gjhs.v8n6p65] [Medline: 26755475]
2. Scaioli G, Schäfer WL, Boerma WG, Spreeuwenberg P, van den Berg M, Schellevis FG, et al. Patients' perception of communication at the interface between primary and secondary care: a cross-sectional survey in 34 countries. BMC Health Serv Res 2019 Dec 30;19(1):1018 [FREE Full text] [doi: 10.1186/s12913-019-4848-9] [Medline: 31888614]
3. van Walraven C, Taljaard M, Bell CM, Etchells E, Zarnke KB, Stiell IG, et al. Information exchange among physicians caring for the same patient in the community. CMAJ 2008 Nov 04;179(10):1013-1018 [FREE Full text] [doi: 10.1503/cmaj.080430] [Medline: 18981442]
4. Stevens DP, Stagg R, Mackay IR. What happens when hospitalized patients see their own records. Ann Intern Med 1977 Apr;86(4):474-477. [doi: 10.7326/0003-4819-86-4-474] [Medline: 300581]
5. Hertz CG, Bernheim JW, Perloff TN. Patient participation in the problem-oriented system: a health care plan. Med Care 1976 Jan;14(1):77-79. [doi: 10.1097/00005650-197601000-00008] [Medline: 1084944]
6. Baldry M, Cheal C, Fisher B, Gillett M, Huet V. Giving patients their own records in general practice: experience of patients and staff. Br Med J (Clin Res Ed) 1986 Mar 01;292(6520):596-598 [FREE Full text] [doi: 10.1136/bmj.292.6520.596] [Medline: 3081187]
7. Golodetz A, Ruess J, Milhous RL. The right to know: giving the patient his medical record. Arch Phys Med Rehabil 1976 Feb;57(2):78-81. [Medline: 1083223]
8. Ross SE, Moore LA, Earnest MA, Wittevrongel L, Lin CT. Providing a web-based online medical record with electronic communication capabilities to patients with congestive heart failure: randomized trial. J Med Internet Res 2004 May 14;6(2):e12 [FREE Full text] [doi: 10.2196/jmir.6.2.e12] [Medline: 15249261]
9. Elkefi S, Yu Z, Asan O. Online medical record nonuse among patients: data analysis study of the 2019 health information national trends survey. J Med Internet Res 2021 Feb 22;23(2):e24767 [FREE Full text] [doi: 10.2196/24767] [Medline: 33616539]
10. Gerber DE, Laccetti AL, Chen B, Yan J, Cai J, Gates S, et al. Predictors and intensity of online access to electronic medical records among patients with cancer. J Oncol Pract 2014 Sep;10(5):e307-e312 [FREE Full text] [doi: 10.1200/JOP.2013.001347] [Medline: 25006222]
11. Patel V, Johnson C. Trends in individuals' access, viewing and use of online medical records and other technology for health needs: 2017-2018. ONC Data Brief. The Office of the National Coordinator for Health Information Technology. 2019 May. URL: https://www.healthit.gov/sites/default/files/page/2019-05/Trends-in-Individuals-Access-Viewing-and-Use-of-Online-Medical-Records-and-Other-Technology-for-Health-Needs-2017-2018.pdf [accessed 2022-11-25]
12. Trivedi N, Patel V, Johnson C, Chou WY. Barriers to accessing online medical records in the United States. Am J Manag Care 2021 Jan;27(1):33-40 [FREE Full text] [doi: 10.37765/ajmc.2021.88575] [Medline: 33471460]
13. Ogink PT, Groot OQ, Karhade AV, Bongers ME, Oner FC, Verlaan JJ, et al. Wide range of applications for machine-learning prediction models in orthopedic surgical outcome: a systematic review. Acta Orthop 2021 Oct;92(5):526-531 [FREE Full text] [doi: 10.1080/17453674.2021.1932928] [Medline: 34109892]
14. Hosmer DWJ, Lemeshow S, Sturdivant RX. Applied Logistic Regression. 3rd edition. Hoboken, NJ, USA: John Wiley & Sons; 2013.

15. Nelder JA, Wedderburn RW. Generalized linear models. J R Stat Soc Ser A 1972;135(3):370-384. [doi: 10.2307/2344614]
16. Breiman L. Random forests. Mach Learn 2001;45:5-32. [doi: 10.1023/A:1010933404324]
17. Rexhepi H, Åhlfeldt RM, Cajander Å, Huvila I. Cancer patients' attitudes and experiences of online access to their electronic medical records: a qualitative study. Health Informatics J 2018 Jun;24(2):115-124 [FREE Full text] [doi: 10.1177/1460458216658778] [Medline: 27440056]
18. Cafri G, Bailey BA. Understanding variable effects from black box prediction: quantifying effects in tree ensembles using partial dependence. J Data Sci 2016;14(1):67-96 [FREE Full text] [doi: 10.6339/jds.201601_14(1).0005]
19. Stewart MT, Hogan TP, Nicklas J, Robinson SA, Purington CM, Miller CJ, et al. The promise of patient portals for individuals living with chronic illness: qualitative study identifying pathways of patient engagement. J Med Internet Res 2020 Jul 17;22(7):e17744 [FREE Full text] [doi: 10.2196/17744] [Medline: 32706679]
20. Anthony DL, Campos-Castillo C, Lim PS. Who isn't using patient portals and why? Evidence and implications from a national sample of US adults. Health Aff (Millwood) 2018 Dec;37(12):1948-1954. [doi: 10.1377/hlthaff.2018.05117] [Medline: 30633673]
21. Hong YA, Jiang S, Liu PL. Use of patient portals of electronic health records remains low from 2014 to 2018: results from a national survey and policy implications. Am J Health Promot 2020 Jul;34(6):677-680. [doi: 10.1177/0890117119900591] [Medline: 32030989]

Abbreviations

AUC: area under the curve
auto-GBM: automatic gradient boosting machine
auto-GLM: automatic generalized linear model
GBM: gradient boosting machine
GLM: generalized linear model
HINTS: Health Information National Trends Survey
LR: logistic regression
OR: odds ratio

Edited by C Lovis; submitted 01.08.22; peer-reviewed by B Ru, J DeShazo, T Ashihara; comments to author 16.11.22; revised version received 26.11.22; accepted 08.04.23; published 19.06.23.

Please cite as:
Chen Y, Liu X, Gao L, Zhu M, Shia BC, Chen M, Ye L, Qin L
Using the H2O Automatic Machine Learning Algorithms to Identify Predictors of Web-Based Medical Record Nonuse Among Patients in a Data-Rich Environment: Mixed Methods Study
JMIR Med Inform 2023;11:e41576
URL: https://medinform.jmir.org/2023/1/e41576
doi:10.2196/41576
PMID:37335616

Original Paper

Exploring Whether the Electronic Optimization of Routine Health Assessments Can Increase Testing for Sexually Transmitted Infections and Provider Acceptability at an Aboriginal Community Controlled Health Service: Mixed Methods Evaluation

Heather McCormack[1,2], MPH; Handan Wand[1], PhD; Christy E Newman[3], PhD; Christopher Bourne[1,2,4], MBBS, MMed; Catherine Kennedy[5], MSc; Rebecca Guy[1], PhD

[1]Kirby Institute, University of New South Wales, Kensington, Australia

[2]Centre for Population Health, New South Wales Ministry of Health, Sydney, Australia

[3]Centre for Social Research in Health, University of New South Wales, Kensington, Australia

[4]Sydney Sexual Health Centre, Sydney, Australia

[5]Maari Ma Health Aboriginal Corporation, Broken Hill, Australia

Corresponding Author:
Heather McCormack, MPH
Kirby Institute
University of New South Wales
Wallace Wurth Building (C27)
High St
Kensington, 2052
Australia
Phone: 61 93481086
Email: hmccormack@kirby.unsw.edu.au

Abstract

Background: In the context of a syphilis outbreak in neighboring states, a multifaceted systems change to increase testing for sexually transmitted infections (STIs) among young Aboriginal people aged 15 to 29 years was implemented at an Aboriginal Community Controlled Health Service (ACCHS) in New South Wales, Australia. The components included electronic medical record prompts and automated pathology test sets to increase STI testing in annual routine health assessments, the credentialing of nurses and Aboriginal health practitioners to conduct STI tests independently, pathology request forms presigned by a physician, and improved data reporting.

Objective: We aimed to determine whether the systems change increased the integration of STI testing into routine health assessments by clinicians between April 2019 and March 2020, the inclusion of syphilis tests in STI testing, and STI testing uptake overall. We also explored the understandings of factors contributing to the acceptability and normalization of the systems change among staff.

Methods: We used a mixed methods design to evaluate the effectiveness and acceptability of the systems change implemented in 2019. We calculated the annual proportion of health assessments that included tests for chlamydia, gonorrhea, and syphilis, as well as an internal control (blood glucose level). We conducted an interrupted time series analysis of quarterly proportions 24 months before and 12 months after the systems change and in-depth semistructured interviews with ACCHS staff using normalization process theory.

Results: Among 2461 patients, the annual proportion of health assessments that included any STI test increased from 16% (38/237) in the first year of the study period to 42.9% (94/219) after the implementation of the systems change. There was an immediate and large increase when the systems change occurred (coefficient=0.22; *P*=.003) with no decline for 12 months thereafter. The increase was greater for male individuals, with no change for the internal control. Qualitative data indicated that nurse- and Aboriginal health practitioner–led testing and presigned pathology forms proved more difficult to normalize than electronic prompts and shortcuts. The interviews identified that staff understood the modifications to have encouraged cultural change around the role of sexual health care in routine practice.

Conclusions: This study provides evidence for the first time that optimizing health assessments electronically is an effective and acceptable strategy to increase and sustain clinician integration and the completeness of STI testing among young Aboriginal people attending an ACCHS. Future strategies should focus on increasing the uptake of health assessments and promote whole-of-service engagement and accountability.

(JMIR Med Inform 2023;11:e51387) doi:10.2196/51387

KEYWORDS

sexual health; sexually transmitted infection; STI; primary care; Indigenous health; electronic medical record; EMR; medical records; electronic health record; EHR; health record; health records; Indigenous; Native; Aboriginal; sexual transmission; sexually transmitted; time series; testing; uptake; acceptance; acceptability; adoption; syphilis; sexually transmitted disease; STD; systems change; health assessment; health assessments; prompt; prompts; implementation; youth; young people; adolescent; adolescents

Introduction

Background

Aboriginal and Torres Strait Islander people in Australia can access sexually transmitted infection (STI) testing at primary health care services, including Aboriginal Community Controlled Health Services (ACCHSs) and mainstream general practice or government sexual health services [1]. The first ACCHS was established in 1971; now, there are >150 ACCHSs nationwide [2]. As in other countries with a history of colonization similar to that of Australia, culturally competent primary health care services run by and for Aboriginal communities play a vital role in ensuring equitable and culturally safe access to sexual and reproductive health care [3-5]. Reflecting this, previous research has identified that ACCHSs are more likely to provide health care that is free of racism than mainstream services [6]. ACCHSs are well situated to embed screening for STIs into routine clinical practice to more regularly identify asymptomatic infections [7].

In Australia, the highest rates of notified STIs (chlamydia, gonorrhea, and syphilis) are among young people aged 15 to 29 years, higher in young Aboriginal and Torres Strait Islander people than in young non-Indigenous people, and the highest in remote areas [8]. *Chlamydia trachomatis* (CT) is the most commonly reported STI among young Aboriginal people, followed by *Neisseria gonorrhea* (NG) and *Treponema pallidum* (syphilis) [9]. CT and NG are easily cured with antibiotics but can be asymptomatic and lead to complications if not treated, such as pelvic inflammatory disease, ectopic pregnancy, and infertility [10-13]. Syphilis is of major concern in Australia, with an expanding epidemic over the last decade beginning in northern Australia and extending to southern Australia, including New South Wales (NSW) and Victoria. Nationally, the notification rate (defined as the rate of notifiable diagnoses per 100,000 person years) is 6 times higher for Aboriginal and Torres Strait Islander people and up to 50 times higher in very remote areas of northern Australia [8]. If left untreated, syphilis can have particularly serious consequences, including miscarriage, stillbirth, and fetal abnormalities [11]. Modeling suggests that the STI prevalence in remote Aboriginal and Torres Strait Islander communities could be considerably reduced if screening was increased to 60% coverage and sustained for 10 years [14], and increasing STI screening rates for young Aboriginal people is a priority target in state and national public health strategies [15,16].

Most ACCHS patients present for reasons other than seeking an STI test; therefore, the mechanisms proposed to increase STI testing in this cohort include integration into other routine primary care visits, such as the Medicare Benefits Schedule item 715 annual health assessment (hereinafter referred to as *health assessment*) [17]. This is a structured preventive health assessment that can be provided annually to Aboriginal and Torres Strait Islander people in primary care settings, including ACCHSs. Australia's universal health insurance system, Medicare, provides a rebate per individual health assessment, and additional practice-level incentives are attached for eligible practices managing patients with chronic disease [18]. The *National guide to a preventive health assessment for Aboriginal and Torres Strait Islander people* [19] recommends the inclusion of STI screening for sexually active young people, along with an assessment of lifestyle factors; social and emotional well-being; vaccination status; and blood lipids, renal function, and blood glucose levels. Approximately one-quarter of young Aboriginal and Torres Strait Islander people participate in a health assessment each year [20]. This is a national figure, and local uptake is highly variable among regions and services [17]. The uptake of health assessments more than tripled in the decade preceding the COVID-19 pandemic, which resulted in a slight decrease of 2% over 2 years [20]. The pandemic also resulted in many health services temporarily deprioritizing the routine STI screening of asymptomatic patients in a context of significant disruption to clinical services [21].

Health assessments involve completing an electronic template in the patient electronic medical record (EMR), which can be customized at the practice level or by the software vendor [22]. The potential of EMR customization to improve the completeness and consistency of health assessments has been anticipated since ACCHSs first began to implement computerized health assessment templates >10 years ago [23]. Electronic prompts have been found to be more effective at increasing STI test requests made by clinicians in other settings than nonelectronic reminders, such as reminder stickers on paper medical records or pathology result forms [24]. The implementation of electronic prompts and automated pathology sets in both the mainstream general practice setting and at government-funded sexual health clinics has been found to increase comprehensive STI testing in line with clinical guidelines and the resulting detection of STIs for patients considered to be at higher risk, such as gay and bisexual men [25-27]. To our knowledge, the use of electronic prompts or

automated pathology sets to increase STI testing in the context of a health assessment at ACCHSs has not been evaluated.

The implementation and integration of complex health care interventions into routine clinical practice relies on action and interaction by and among individuals—the "work" of normalizing and embedding change [28]. In this paper, normalization is defined as per May et al [29] as the work of actors when engaging with new or changed activities by which the new activities become routinely embedded into existing knowledge and practices. Through this lens, implementation is characterized as an inherently social process involving collective action from human actors within dynamic contexts [30]. This provides a framework to analyze the factors that encourage or inhibit the incorporation of health care interventions into routine clinical practice [31].

Objectives

This mixed methods evaluation examined whether the systems change implemented at an NSW ACCHS increased the integration of STI testing into health assessments in young people, with a particular focus on concurrent syphilis testing, and explored the understandings of the acceptability and normalization of the systems change among ACCHS staff.

Methods

Setting

This study was conducted in partnership with a western NSW ACCHS caring for a large Aboriginal population [32] of between 8.4% in the main town center and 61.2% in smaller satellite towns (compared with 2.9% for NSW as a whole) [33,34]. This ACCHS provides 53,000 occasions of service to an average of 3500 patients per year, 80% of whom are Aboriginal patients. This figure accounts for all separate interactions with patients and includes face-to-face visits and telephone calls, both clinical and administrative. As the Torres Strait Islander population is low, we will hereinafter respectfully use the term *Aboriginal* when referring to this setting. Although NSW is home to the largest Aboriginal population in Australia, there is little quantitative sexual health research conducted with Aboriginal populations in NSW [35], although some qualitative studies have been published in recent years [36,37].

The Systems Change

We collaboratively developed a multifaceted systems change with ACCHS staff in March 2019. ACCHS clinical, managerial, and administrative staff who had been identified as potential sexual health champions cocreated the systems change at a workshop facilitated by 2 members of the research team (HM and CB). The workshop aimed to identify strategies to improve syphilis testing rates and STI testing rates more broadly among the patient population within the existing range of services, activities, and staff skills available. A core component was the addition to the EMR of an electronic prompt and an automated pathology set. The prompt was attached to the health assessment template completed by general practitioners (GPs) and triggered for all patients in the target age range, advising the GP to query sexual activity since the patient's last test because guidelines recommend STI screening after a change in partner or otherwise annually [38]. An affirmative indication updated the health assessment template in the EMR and autopopulated a pathology request for a comprehensive STI test (chlamydia, gonorrhea, and syphilis). The automated pathology set replaced the manual completion of pathology requests for individual STI tests with a single-click button that autopopulated a pathology request for a comprehensive STI test. The prompt aimed to increase the integration of STI tests into the health assessment, and the automated pathology test set aimed to increase the inclusion of syphilis tests when STI tests were requested, both within and outside of the health assessment.

Three other operational components completed the package: credentialing registered nurses (RNs) and Aboriginal health practitioners (AHPs) to undertake STI screening, pathology request forms presigned by a GP to authorize tests conducted by RNs and AHPs [39], and the enhanced internal reporting of STI screening targets to staff via standing agenda items at staff meetings for detailed data reports. Credentialing RNs and AHPs aimed to create opportunities for STI testing in clinical encounters that did not involve a GP, whereas having pathology forms presigned by a GP allowed these tests to receive Medicare funding under the signing GP's provider number. Enhanced data reporting has been shown to be an important component of continuous quality improvement (CQI) systems implemented at ACCHSs [40,41]. These 3 components aimed to increase STI testing outside of the health assessment.

Strengths-Based Approach

Our evaluation explicitly adopted a strengths-based approach. This project centers the strengths, assets, and capabilities of ACCHSs and ACCHS staff [42], rather than deficit-focused epidemiological framing that can act as a barrier to improving health outcomes [43]. As such, both the qualitative and quantitative analyses focused on factors associated with positive health outcomes for Aboriginal people rather than measuring ill health against a non-Indigenous baseline [44], and the integration of findings in the discussion foregrounded the strengths and enablers of success rather than barriers and deficiencies.

Mixed Methods Approach

Our mixed methods evaluation used an explanatory sequential design as defined by Creswell and Clark [45], with quantitative data collected first and then used to inform qualitative data collection and analysis. This design supported the use of in-depth qualitative analysis to more closely examine potential explanations for the quantitative findings, which provides greater insight into the performance of the systems change than quantitative analysis alone [45]. The interview guide for the qualitative component was informed by preliminary findings from the quantitative component, with the qualitative interviews then exploring how staff understood the implementation and outcomes of the systems change. Our qualitative analysis allowed the researchers to explore how and why the components of the systems change had or had not been successfully normalized into staff experiences of routine practice, building on our quantitative analysis to provide a more whole and complex picture of the success of the systems change. The

quantitative and qualitative findings have been narratively integrated and explored together in the *Discussion* section.

Quantitative Methods

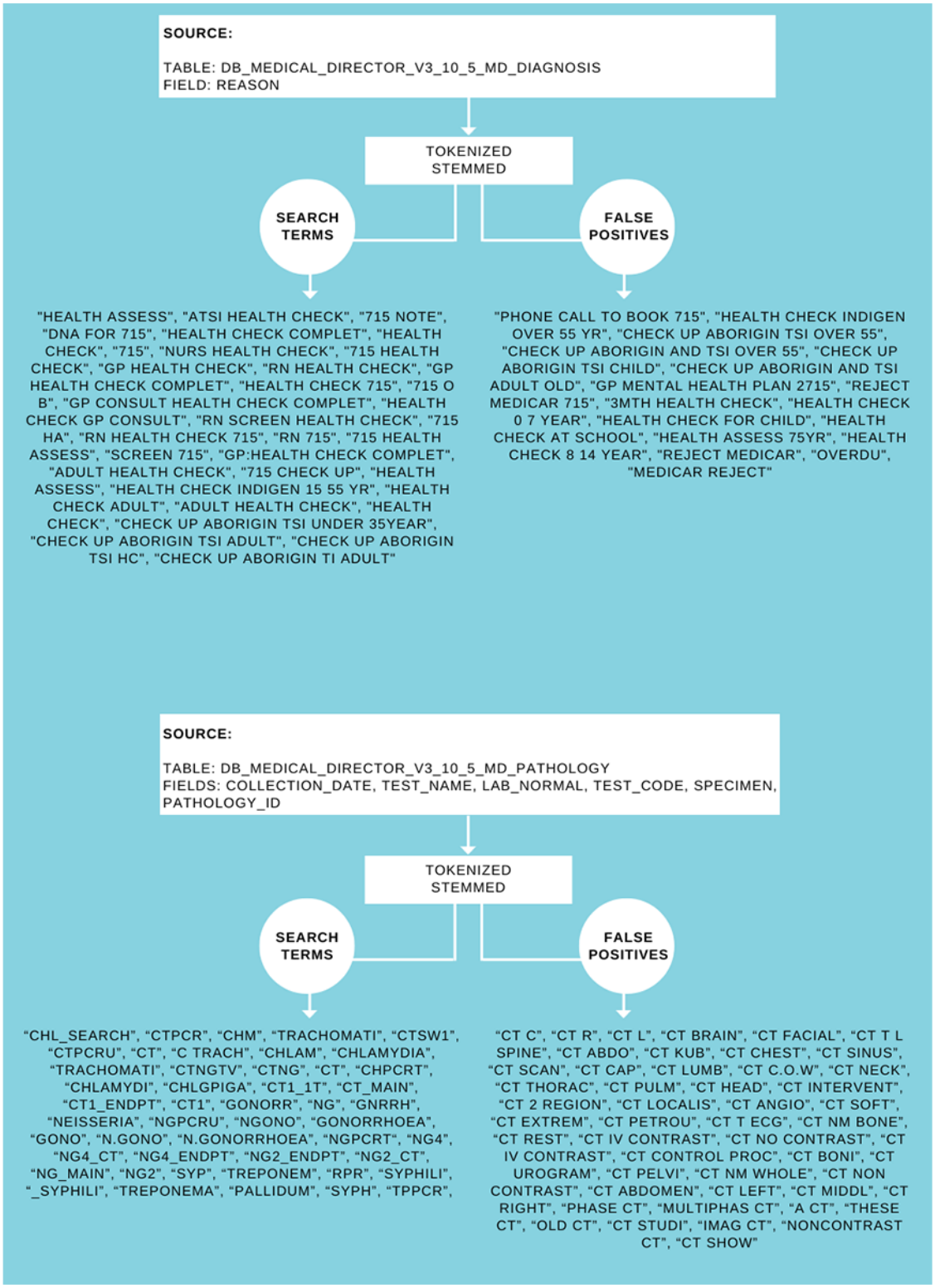
Our analysis was based on an annually deduplicated data set of unique Aboriginal patients aged 15 to 29 years who attended the NSW ACCHS in the 24 months before implementation and in the 12 months after implementation (April 2017-March 2020). Using the exact probabilities of the binominal distribution, we determined that a sample size of ≥2000 patients would be sufficient to detect a difference of 5% to 7% in the proportion of those who had an STI test between the intervention and control periods, assuming that STI testing coverage was 10% to 20%, with 80% power.

The primary outcome for the quantitative component of this study was the proportion of health assessments that included a test for any STI (chlamydia, gonorrhea, or syphilis). The secondary outcomes were the proportion of health assessments that included all 3 STI tests, the proportion of health assessments inclusive of a chlamydia and gonorrhea test that also included a syphilis test (concurrent testing), the proportion of the total study population that received a test for any STI (chlamydia, gonorrhea, or syphilis), and the proportion of the total study population that received a health assessment. The control outcome was the proportion of health assessments that included a random blood glucose level test, a recommended inclusion in the health assessment unaffected by the systems change.

Under strict confidentiality and privacy protocols, deidentified routinely collected clinical data for all young Aboriginal people aged 15 to 29 years who attended the ACCHS in the study period (April 2017-March 2020) were extracted from the EMR using software called GRHANITE (Generic Health Network Information Technology for the Enterprise) [46] via the enrolment of the service in the ATLAS sentinel surveillance network, which has been described elsewhere [47]. We estimated the average number of clinical interactions per person and their 95% CIs using Poisson regression. Demographic variables included age, sex, and Aboriginal status. Testing variables collected via ATLAS included the test performed (chlamydia, gonorrhea, or syphilis) and the date the test was conducted. A further testing variable (random blood glucose level test) was collected manually by the in-house ACCHS statistician because it was outside the scope of ATLAS. ATLAS data collected via natural language processing and keyword search was validated with the help of ACCHS clinical staff. The ATLAS keywords are shown in Figure 1. As tests were conducted at the time of consultation at an in-house pathology service, the date of the test request is not reported separately. Attendance and visit variables included the date of consultation and Medicare Benefits Schedule item for health assessment.

Annual changes in each outcome were examined by age group and sex. An interrupted time series (ITS) regression model (also known as segmented regression) was used to estimate the changes in the proportion of any STI test conducted among those who had a health assessment over time. Regression coefficients with 95% CIs were presented for comparing the before and after periods. An ITS is a technique used for evaluating what are sometimes referred to as natural experiments by estimating the consequence of specific changes in a real-world setting and adjusting for individual- and community-level factors via multilevel regression models to account for covariates [48]. The ITS analysis was then stratified by sex for the primary outcome and the control outcome.

Figure 1. Health assessment and sexually transmitted infection test request data extract coding.



Qualitative Methods

The staff of the NSW ACCHS were the population for the qualitative component of this study. Stratified purposeful sampling [49] was used to recruit staff, who were informed of the project at an in-service information program delivered by members of the research team (HM and CB) and then invited to participate via email, with key staff members also contacted by telephone. Recruitment remained open until the sample contained representation from each of the 3 inclusion categories of clinical staff (GPs, RNs, and AHPs). Because of the impacts of the COVID-19 pandemic, interviews were not completed until the year after the implementation of the systems change. Each participant completed a semistructured interview lasting approximately 1 hour that followed an interview guide informed by normalization process theory (NPT) [28,30,50]. Previous sexual health research has drawn on NPT to examine provider-initiated HIV testing and counseling in South Africa

[51] and annual STI screening in remote ACCHSs in the Northern Territory, Australia [52,53]. As well as open-ended questions on topics such as how the systems change affected work practices, the factors that made the components of the systems change easier or more difficult to implement, perceived impacts on the service, and suggestions for future modifications, the interviews included questions informed by the quantitative findings. Participants were asked whether they thought that STI testing rates had increased after the systems change, then presented with preliminary quantitative data and asked for their opinions on how the components of the systems change had contributed to these findings.

The interviews were transcribed and analyzed thematically by the first author (HM) as per Braun and Clarke [54] and later works [55,56] using NVivo software (Lumivero). HM is an Aboriginal woman from NSW with expertise in sexual health. She was supported throughout the analysis by CN, a non-Indigenous researcher with substantial experience in qualitative approaches to sexual health research. Codes were informed by NPT [31,50], focusing on the 4 main constructs of coherence, cognitive participation, collective action, and reflexive monitoring. Component definitions within each construct were informed by the study conducted by Hengel et al [53], which used NPT to assess the integration into routine practice of an ACCHS CQI program. Textbox 1 (adapted from Hengel et al [53] and Murray et al [31]) outlines NPT constructs and components.

Textbox 1. Normalization process theory constructs and components.

Construct, component, and definition of component

- Coherence
  - Differentiation: perceived difference from existing practice
  - Communal specification: shared understanding of the aims of the systems change
  - Individual specification: individual understanding of the aims of the systems change
  - Internalization: perceived importance and value of the systems change
- Cognitive participation
  - Initiation: willingness to drive the systems change forward
  - Enrolment: collaborative work to make the systems change succeed
  - Legitimation: perceptions of the worthiness of time and effort
  - Activation: likelihood of participants sustaining the changes to work practice
- Collective action
  - Interactional workability: effect of the systems change on shared work
  - Relational integration: accountability built on knowledge
  - Skill set workability: appropriateness of the systems change to participant skill sets
  - Contextual integration: compatibility with existing policies, practices, and environmental context
- Reflexive monitoring
  - Systematization: effectiveness of the systems change in daily practice
  - Communal appraisal: group evaluation of the systems change
  - Individual appraisal: personal relationship with the systems change
  - Reconfiguration: opportunities for participant experience to inform modifications to systems change

HM developed a thematic map and undertook a reread of the raw data to recode any data that were missed in the earlier stages of analysis. Codes, themes, and the analytic framework were discussed by all authors throughout the coding and analysis process to make sense of shared understandings and to interrogate assumptions and interpretations.

Ethical Considerations

The evaluation was overseen by a research governance committee with representation from the medical staff, nursing staff, public health staff, and primary health care management staff of the ACCHS. An ACCHS staff member also contributed to the research team. This study received ethics approval through the ethics committee of the Aboriginal Health & Medical Research Council of NSW (HREC 1700/20). Participants were informed both at the in-service program and before each interview that their participation was voluntary and provided with a participation information sheet and consent form, the contents of which were read aloud at the commencement of the interview. Informed voluntary consent was obtained before the collection of any personal information. The research team includes 2 researchers with an existing relationship with the

health service with no authoritative influence over prospective interview participants.

Results

Quantitative Results

During the study period (April 2017-March 2020), a total of 54,299 patient interactions were recorded in the EMR; after excluding interactions recorded with non-Indigenous patients (n=6317, 11.63%) and deduplication, 2461 (4.53%) unique Aboriginal patients who attended the ACCHS for a clinical interaction at least once, with an average of 19 interactions (SE 0.09; 95% CI 19.32-19.68) recorded in the EMR per patient, were included in the analysis. The recorded interactions per patient may have occurred at a single visit or across multiple visits. Among these 2461 unique Aboriginal patients, 644 (26.17%) had at least 1 episode of a health assessment annually recorded as a clinical interaction in the EMR. Of the 2461 unique Aboriginal patients, 45.43% (n=1118) were men, and 54.57% (n=1343) were women, with a median age of 22 (IQR 18-26) years. The study population is shown in Figure 2.

Overall, 27.2% (175/644) of the health assessments included any STI test, 23.6% (152/644) included a CT/NG test, and 18.5% (119/644) included a syphilis test. Combined CT/NG and syphilis testing occurred in 14.9% (96/644) of the health assessments.

Compared with the 24 months before implementation, the annual proportion of health assessments that integrated any STI test (CT/NG and syphilis) increased in the postintervention period (year 1: 38/237, 16%; year 2: 43/188, 22.9%; and intervention period: 94/219, 42.9%). Greater increases were observed in those aged 20 to 24 years and 25 to 29 years than in those aged 15 to 19 years. The increase by age group was similar for all study outcomes. The changes in the integration of a CT/NG test into a health assessment were similar for both male individuals and female individuals, although male individuals had a higher baseline value. However, the increase in concurrent syphilis testing with CT/NG testing was more substantial for male individuals than for female individuals. Similar results by patient sex were observed for the integration of any STI test and the integration of both a CT/NG test and a syphilis test (Table 1).

Of the health assessments that included a CT/NG test, 63.2% (96/152) overall also included a concurrent syphilis test.

Figure 2. Study population.

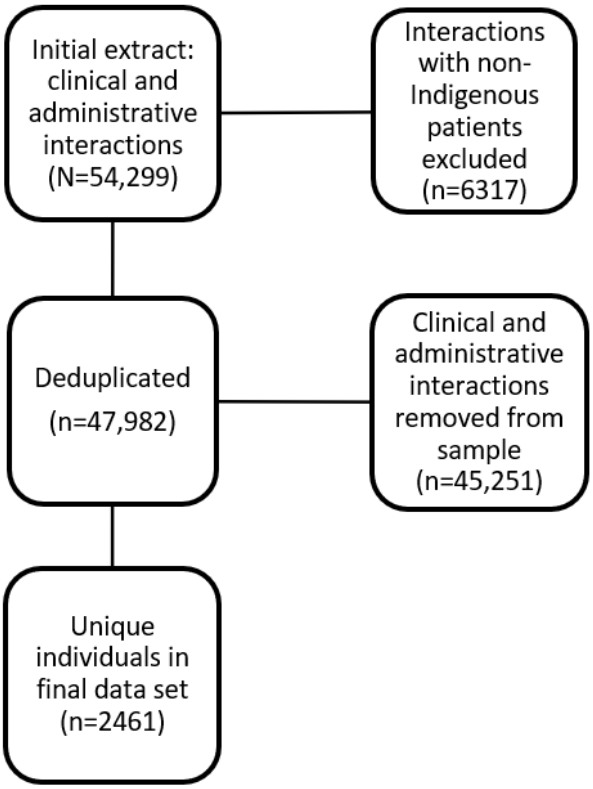


Table 1. Proportion of health assessments conducted on young Aboriginal people aged 15 to 29 years during the study period that included tests for *Chlamydia trachomatis* (CT) and *Neisseria gonorrhea* (NG) or syphilis (n=644).

Time period and breakdown	Any STI[a] test included (CT or NG or syphilis; n=175), n (%)	CT and NG test included (n=152), n (%)	Syphilis test included (n=119), n (%)	Combined CT and NG and syphilis testing included (n=96), n (%)
April 2017-March 2018 (year 1; n=237)				
Breakdown by age group (years)				
15-19 (n=86)	14 (16.3)	12 (14)	6 (7)	4 (4.7)
20-24 (n=81)	13 (16.1)	10 (12.4)	6 (7.4)	3 (3.7)
25-29 (n=70)	11 (15.7)	10 (14.3)	7 (10)	6 (8.6)
Breakdown by sex				
Male (n=119)	23 (19.3)	21 (17.7)	9 (7.6)	7 (5.9)
Female (n=118)	15 (12.7)	11 (9.3)	10 (8.5)	6 (5.1)
April 2018-March 2019 (year 2; n=188)				
Breakdown by age group (years)				
15-19 (n=67)	13 (19.4)	12 (17.9)	5 (7.5)	4 (6)
20-24 (n=63)	17 (27)	16 (25.4)	12 (19.1)	11 (17.5)
25-29 (n=58)	13 (22.4)	11 (19)	9 (15.5)	7 (12.1)
Breakdown by sex				
Male (n=89)	22 (24.7)	21 (23.6)	14 (15.7)	13 (14.6)
Female (n=99)	21 (21.2)	18 (18.2)	12 (12.1)	9 (9.1)
April 2019-March 2020 (intervention; n=219)				
Breakdown by age group (years)				
15-19 (n=79)	24 (30.4)	20 (25.3)	18 (22.8)	14 (17.7)
20-24 (n=69)	34 (49.3)	32 (46.4)	25 (36.2)	23 (33.3)
25-29 (n=71)	36 (50.7)	29 (40.9)	31 (43.7)	24 (33.8)
Breakdown by sex				
Male (n=113)	55 (48.7)	49 (43.4)	49 (43.4)	43 (38.1)
Female (n=106)	39 (36.8)	32 (30.2)	25 (23.6)	18 (17)

[a]STI: sexually transmitted infection.

The annual proportion of health assessments inclusive of a CT/NG test that also included a syphilis test increased in all age groups, although those aged 25 to 29 years had higher baseline values than those aged 15 to 19 years or 20 to 24 years (Table 2). The inclusion of a syphilis test in health assessments with a CT/NG test increased for male individuals to 88% (18/32) but remained relatively stable for female individuals.

The annual proportion of young people who received a health assessment fluctuated slightly, with uptake slightly higher in male individuals than in female individuals and in those aged 15 to 19 years than in older young people throughout the study period. Among unique young Aboriginal people attending the service each year, there was a small increase in the proportion tested for CT/NG (from 195/800, 24.4% to 243/823, 29.5%) and a greater increase in the proportion tested for syphilis (from 103/800, 12.9% to 225/823, 27.3%), with increases greater for male individuals than for female individuals for both as well as the outcome: receiving all STI tests (Table 3).

The ITS model shown in Figure 3 shows a significant increase in the quarterly proportion of health assessments completed with young Aboriginal people that included an STI test immediately after implementation (coefficient=0.22; *P*=.003), greater among male individuals (coefficient=0.28; *P*=.01) than among female individuals (coefficient=0.15; *P*=.005; Table 4; Figures 4 and 5). There was no trend in the quarterly intervals in the postintervention period for either male or female individuals.

When the ITS analysis was repeated with the final quarter excluded (January-March 2020, the beginning of the COVID-19 pandemic), an increasing trend was observed (*P*=.006, data not shown).

There was no immediate change in the quarterly proportion of health assessments that included a random blood glucose level test (Table 5), overall (Figure 6) and for male individuals (Figure 7) and female individuals (Figure 8). In the postintervention period, no trend was detected overall (*P*=.90) or in male individuals (*P*=.20) and female individuals (*P*=.20).

Table 2. Proportion of health assessments conducted in young Aboriginal people aged 15 to 29 years inclusive of a *Chlamydia trachomatis* and *Neisseria gonorrhea* (CT/NG) test that concurrently included a syphilis test (n=152).

Time period and breakdown	Health assessments with a CT and NG test that concurrently also included a syphilis test, n (%)
April 2017-March 2018 (year 1; n=32)	
Breakdown by age group (years)	
15-19 (n=12)	4 (33)
20-24 (n=10)	3 (30)
25-29 (n=10)	6 (60)
Breakdown by sex	
Male (n=21)	7 (33)
Female (n=11)	6 (55)
April 2018-March 2019 (year 2; n=39)	
Breakdown by age group (years)	
15-19 (n=12)	4 (33)
20-24 (n=16)	11 (69)
25-29 (n=11)	7 (64)
Breakdown by sex	
Male (n=21)	13 (62)
Female (n=18)	9 (50)
April 2019-March 2020 (intervention; n=81)	
Breakdown by age group (years)	
15-19 (n=20)	14 (70)
20-24 (n= 32)	23 (72)
25-29 (n=29)	24 (83)
Breakdown by sex	
Male (n=49)	43 (88)
Female (n=32)	18 (56)

Table 3. Proportion of young Aboriginal people aged 15 to 29 years attending the service during the study period who received a sexually transmitted infection (STI) test (chlamydia, gonorrhea, or syphilis) or a health assessment (n=2461).

Time period and breakdown	Received a CT[a] and NG[b] test (n=637), n (%)	Received a syphilis test (n=470), n (%)	Received all STI tests (n=330), n (%)	Received a health assessment (n=644), n (%)
April 2017-March 2018 (year 1; n=800)				
Breakdown by age group (years)				
15-19 (n=276)	48 (17.4)	21 (7.6)	14 (5.1)	86 (31.2)
20-24 (n=273)	74 (27.1)	40 (14.7)	30 (11)	81 (29.7)
25-29 (n=251)	73 (29.1)	42 (16.7)	35 (13.9)	70 (27.9)
Breakdown by sex				
Male (n=364)	65 (17.9)	38 (10.4)	32 (8.8)	119 (32.7)
Female (n=436)	130 (29.8)	65 (14.9)	47 (10.8)	118 (27.1)
April 2018-March 2019 (year 2; n=838)				
Breakdown by age group (years)				
15-19 (n=285)	515 (17.9)	295 (10.2)	18 (6.3)	67 (23.5)
20-24 (n=280)	79 (28.2)	60 (21.4)	42 (15)	63 (22.5)
25-29 (n=273)	69 (25.3)	53 (19.4)	32 (11.7)	58 (21.3)
Breakdown by sex				
Male (n=384)	70 (18.2)	43 (11.2)	33 (8.6)	89 (23.2)
Female (n=454)	129 (28.4)	99 (21.8)	59 (13)	99 (21.8)
April 2019-March 2020 (intervention; n=823)				
Breakdown by age group (years)				
15-19 (n=279)	59 (21.2)	50 (17.9)	32 (11.5)	79 (28.3)
20-24 (n=280)	95 (33.9)	83 (29.6)	61 (21.8)	69 (24.6)
25-29 (n=264)	89 (33.7)	92 (34.9)	66 (25)	71 (26.9)
Breakdown by sex				
Male (n=370)	98 (26.5)	94 (25.4)	77 (20.8)	113 (30.5)
Female (n=453)	145 (32)	131 (28.9)	82 (18.1)	106 (23.4)

[a]CT: *Chlamydia trachomatis.*

[b]NG: *Neisseria gonorrhea*.

Figure 3. Interrupted time series analysis of health assessments inclusive of any sexually transmitted infection (STI) test (chlamydia, gonorrhea or syphilis): overall study population. Regression with Newey-West SEs: lag(1).

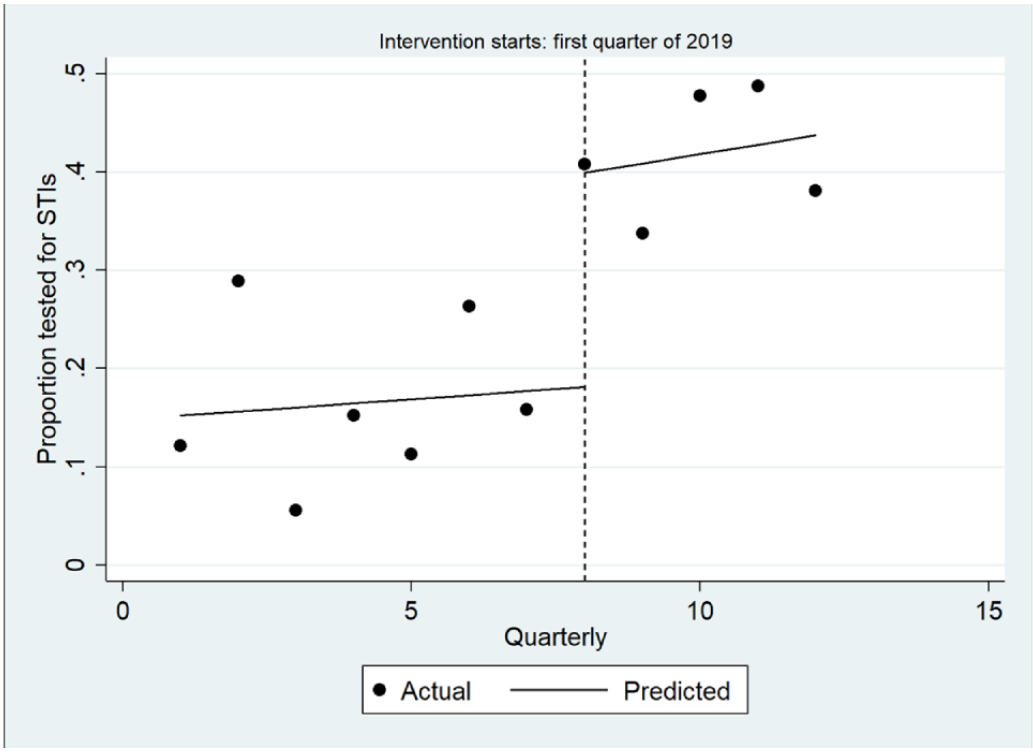


Table 4. Estimates from the interrupted time series analysis for the proportion of health assessments inclusive of any sexually transmitted infection test (chlamydia, gonorrhea, or syphilis).

	Overall		Male		Female	
	Coefficient (95% CI)	*P* value	Coefficient (95% CI)	*P* value	Coefficient (95% CI)	*P* value
Before the intervention period of March 2019	0.004 (−0.02 to 0.03)	.70	−0.003 (−0.04 to 0.03)	.08	0.01 (−0.008 to 0.030)	.18
Immediate effect of the intervention period of March 2019	0.22 (0.10 to 0.34)	.003	0.28 (0.08 to 0.50)	.01	0.15 (0.06 to 0.24)	.005
Overall effect of the intervention period of March 2019	0.01 (−0.02 to 0.04)	.40	0.01 (−0.02 to 0.04)	.40	0.01 (−0.04 to 0.06)	.67

Figure 4. Interrupted time series analysis of health assessments inclusive of any sexually transmitted infection (STI) test (chlamydia, gonorrhea or syphilis): male. Regression with Newey-West SEs: lag(1).

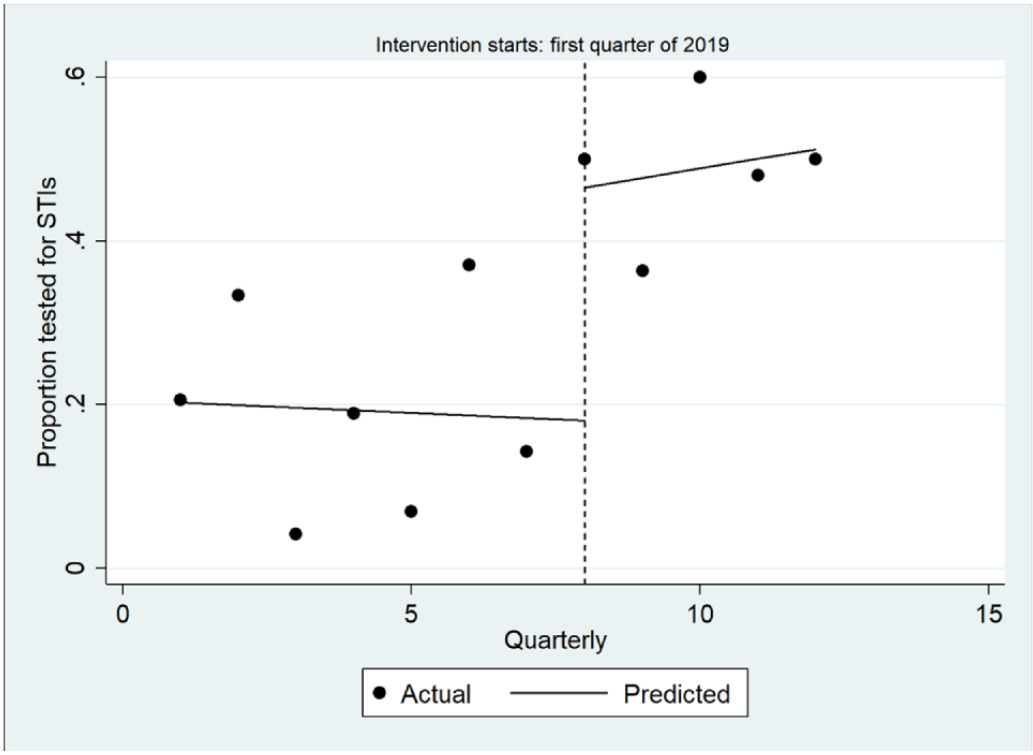


Figure 5. Interrupted time series analysis of health assessments inclusive of any sexually transmitted infection (STI) test (chlamydia, gonorrhea or syphilis): female. Regression with Newey-West SEs: lag(1).

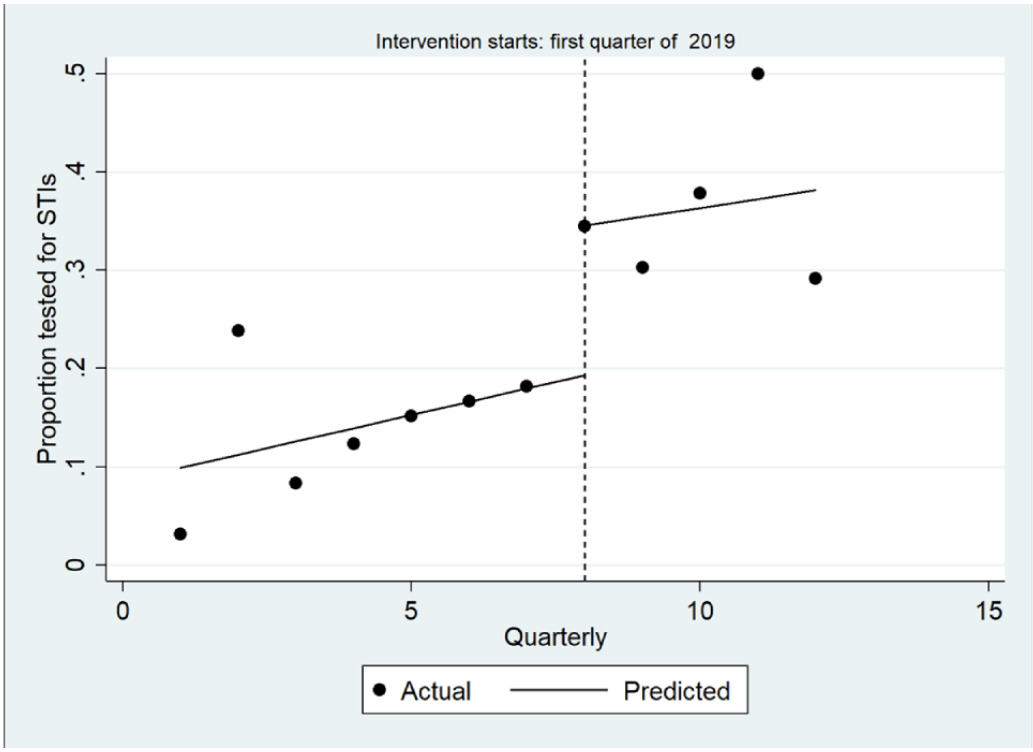


Table 5. Estimates from the interrupted time series for the proportion of health assessments that included a random blood glucose level test.

	Overall		Male		Female	
	Coefficient (95% CI)	*P* value	Coefficient (95% CI)	*P* value	Coefficient (95% CI)	*P* value
Before the intervention period of March 2019	0.01 (−0.003 to 0.030)	.10	0.02 (−0.004 to 0.050)	.09	0.002 (−0.03 to 0.03)	.90
Immediate effect of the intervention period of March 2019	−0.02 (−0.11 to 0.07)	.55	−0.09 (−0.31 to 0.14)	.40	0.01 (−0.18 to 0.20)	.90
Overall effect of the intervention period of March 2019	−0.002 (−0.03 to 0.03)	.90	0.05 (−0.03 to 0.12)	.22	−0.03 (−0.09 to 0.02)	.20

Figure 6. Interrupted time series analysis of health assessments inclusive of a random blood glucose level test: overall study population. Regression with Newey-West SEs: lag(1).

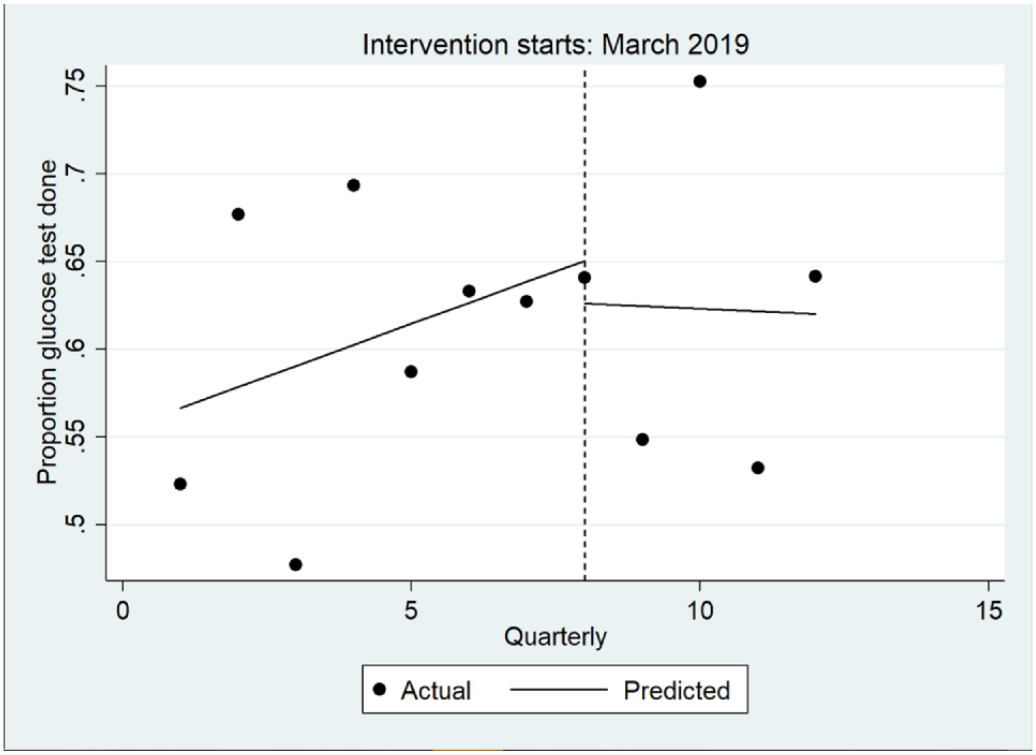


Figure 7. Interrupted time series analysis of health assessments inclusive of a random blood glucose level test: male. Regression with Newey-West SEs: lag(1).

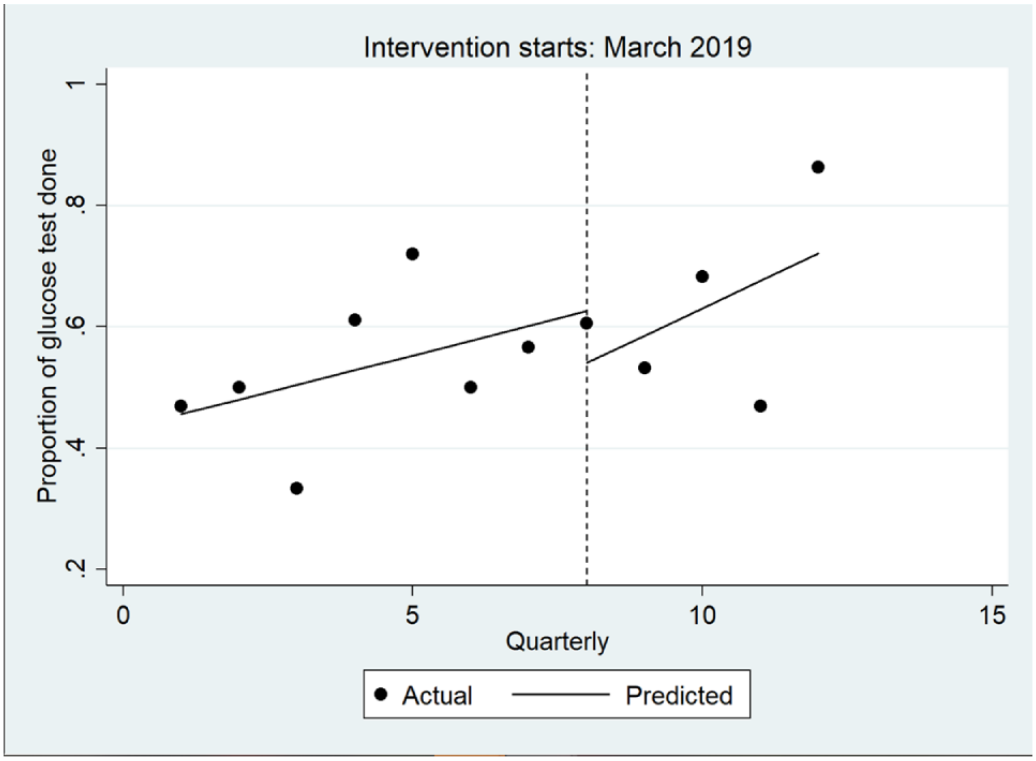
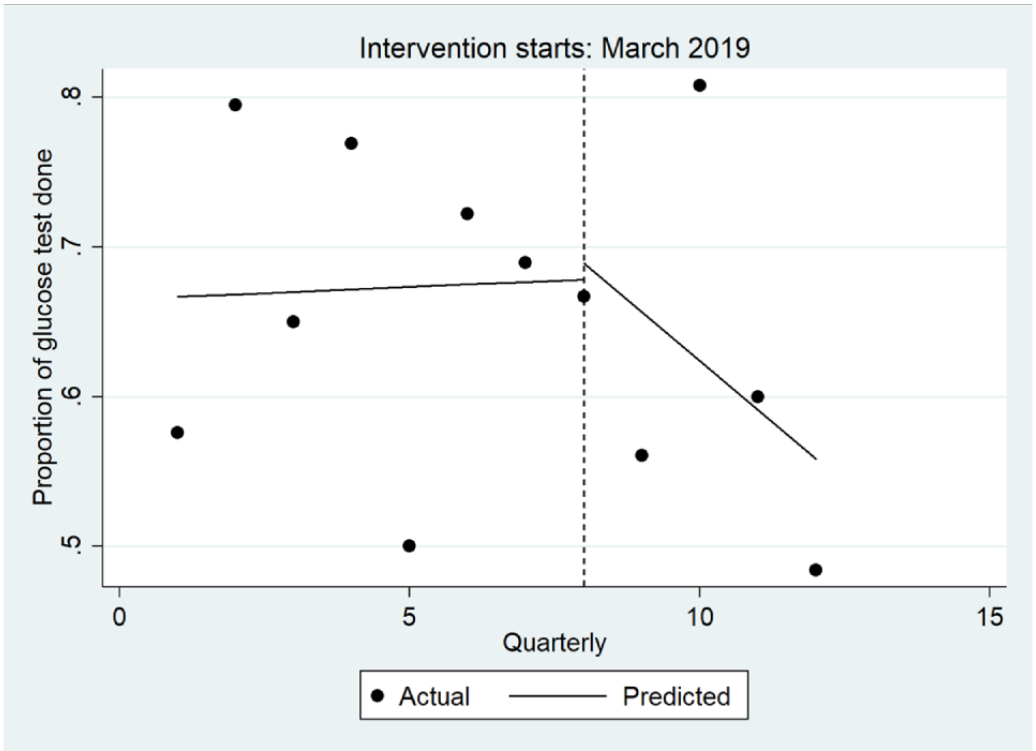


Figure 8. Interrupted time series analysis of health assessments inclusive of a random blood glucose level test: female. Regression with Newey-West SEs: lag(1).



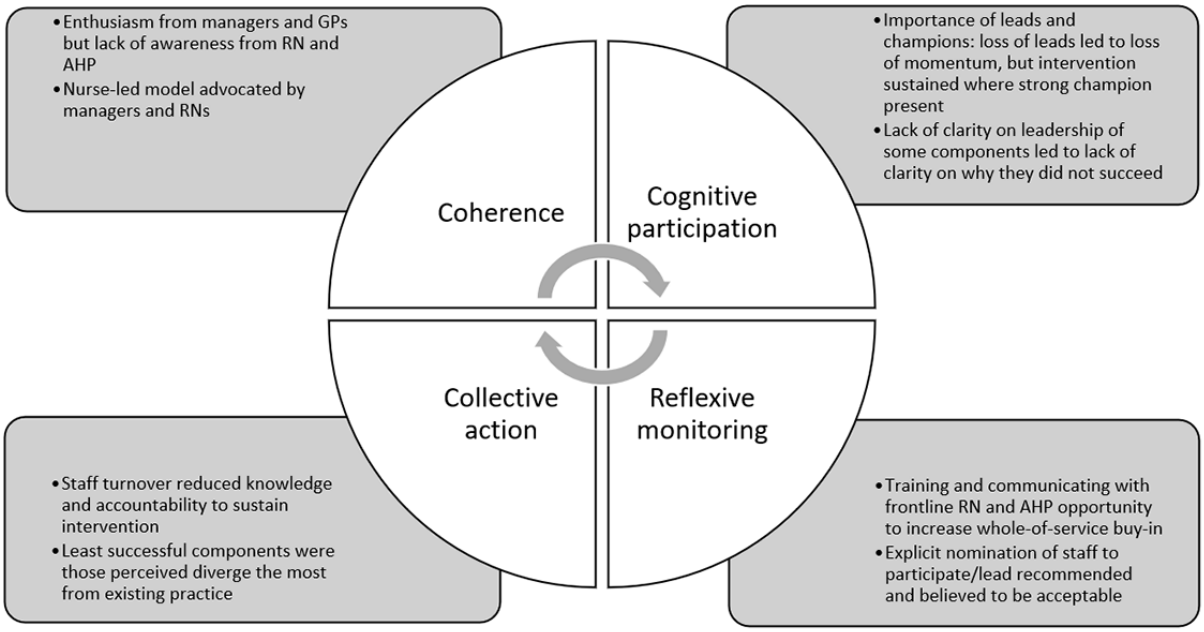
Qualitative Results

Overview

A total of 6 ACCHS staff members participated in an interview: managers (n=2, 33%), GPs (n=2, 33%), an RN (n=1, 17%), and an AHP (n=1, 17%), representing a little more than half of the staff members who had maintained continuous employment from before the implementation until after the end of the postintervention period. The qualitative interviews explored potential explanations for the quantitative findings through seeking to understand how ACCHS staff had experienced and perceived the implementation of the systems change and its impact both on the STI testing rates examined in the quantitative

analysis and on everyday work practice. This allowed the research team to tease out which components of the systems change had or had not been successfully normalized, what factors had contributed to normalization, and how the successful or unsuccessful normalization of different components of the systems change explained key findings from the quantitative analysis. The coding framework consists of 4 NPT constructs: coherence (how participants make sense of the change), cognitive participation (engagement by participants), collective action (the participants' work that has contributed to the intervention functioning), and reflexive monitoring (participant assessment of benefits and costs). Figure 9 captures the key insights from the qualitative data of relevance to each of the 4 NPT constructs.

Figure 9. Thematic map of qualitative analysis informed by normalization process theory. AHP: Aboriginal health practitioner; GP: general practitioner; RN: registered nurse.



Coherence

Managers and GPs displayed enthusiasm about the electronic prompts and automated pathology sets, but the RN and the AHP, whose roles did not require them to use the prompts and test sets, demonstrated low awareness. Communication to frontline staff about the systems change seemed to have been inadequate to instill a shared understanding of its importance and value. However, participants who demonstrated a high awareness of the systems change identified that the systematized incorporation of age-based STI testing into the health assessment had had a broader impact on organizational norms around sexual health in routine practice, with several participants identifying this as a potential explanation for the observed increase in STI testing:

> *Previously it was an individual clinician who may or may not have decided this patient should be tested. Like there was no organizational-wide policy or...expectation that clinicians would be doing that.... Our [health assessments] were changed to include the sexual health testing for people in the targeted age range. So that lifted sexual health screening from sort of some nebulous thing that people may or may not do...to something that was very much at the forefront and was discussed at clinical meetings. It really gained a life of its own which it had never had previously.* [Participant 1, manager]

Participants expressed appreciation for RNs and AHPs, but the RN and AHP participants felt that non-GP clinical roles were not leveraged to their full potential and that this had limited the successful implementation of the systems change. All participants recognized the importance of the "soft" skill sets of RNs and AHPs in patient engagement. Adding to this with enhanced clinical autonomy and a nurse-led model of care via credentialing RNs and AHPs to complete STI screens independently was broadly supported by staff at all levels. Participants considered the nurse credentialing component of the package and the presigned pathology forms component to be strongly interlinked. However, whereas participants reported strong support for the credentialing component, the pathology forms met resistance among the GP workforce, although the RN participant was supportive of the component:

> *One thing I encounter when I was working here is things like, just ordering pathology form, like STI check, for example, if they can allow the nurses to do that, you know. Because sometimes people just come in.... They are asymptomatic.... But they don't want to see the doctor...if I'm given the autonomy to practice, I could just give the, print out a pathology form and order it.... If the nurses are allowed to do, to order things like that, STI check, then it may improve the STI screening.* [Participant 6, RN]

The underuse of the non-GP personnel was also apparent in participant discussions of data reporting. An increased

transparency of data was perceived positively, and several participants described an increased awareness of trends in practice data as motivating for the workforce. However, GPs and managers again had much greater awareness of the changes that had been made than the RN and AHP participants, although the RN and AHP participants felt that strengthening communication with frontline staff would increase whole-of-service commitment to improving outcomes. Participants did not identify any linkages between the improvements made to data analysis capacity and the quantitative findings.

Cognitive Participation

All participants perceived leads and champions as vital to moving the systems change forward, which was most evident in relation to the electronic prompts and automated pathology sets. Strong leadership encouraged early adoption among the GP workforce, but the departure of people in key lead or champion roles led to this action being deprioritized among some newer GPs. It was sustained consistently among GPs who had been present before the development of the package, but only implemented ad hoc for newer GPs after the departure of a key GP champion partway through the study period. Participants perceived staff turnover as a barrier to normalizing this component and felt that the positive impact seen on STI testing rates within the health assessment was the result of the commitment of longer-term staff to this component of the systems change.

Participants were unclear who held responsibility for the credentialing of RNs and AHPs. Some RNs completed training, but this lack of clarity impeded the necessary systemic change to routine practice. The presigned pathology forms were successfully sustained in the prenatal setting owing to a strong GP champion, and there was interest from RNs service-wide, but outside of the prenatal setting, no individual GP had self-nominated as the signatory. The lack of GP commitment to this component also prevented the implementation of the credentialing component because RNs who had completed credentialing could not order tests without presigned forms. As the aim of these components was to enable RN- and AHP-led testing outside of GP consultations, this explains the lower increase in STI testing outside of the health assessment:

> *I think a GP champion and I'm gonna say the sexual health physician that should be full time in that sexual health area.... Apart from the fact we have an inexperienced, untrained person in the job, there's just not enough time left over to give it, to actually do all that; you know, that working one-on-one with people and following up, and talking to the GPs, and...yeah. So that's what I'd say. A GP champion and a full-time sexual health RN or health worker.* [Participant 1, manager]

However, participants did not always expect champions to be senior clinicians or managers. The standardization of the health assessment was perceived to provide systemic support for AHPs to champion the inclusion of routine STI testing to GPs who may otherwise use their discretion to not include it, which further supports the systems change as the driver of increased testing within the health assessment. Participants expressed that this normalization of practice had partially addressed the concern regarding some newer or locum GPs not having implemented the EMR modifications:

> *If we've got good-quality, well-trained health workers, I think that's a really big plus and a really good way for it to go forward, 'cause it doesn't depend on the GP who might be a locum for 2 weeks: it depends on that health worker going, "Part of our [health assessment] is that we always do this comprehensive STI screen.*" [Participant 4, GP]

Although participants supported improvements in data reporting in principle, they almost universally perceived it as something that they were recipients of, rather than participants in. The interviews depicted an understanding of data reporting as a contribution to the service made by an unidentified other that would then inform participants' work, rather than a reflection of the outcomes of their work that could meaningfully contribute to quality improvement. This perhaps explains why participants did not perceive a strong relationship between this component and the quantitative findings.

Collective Action

Considerable staff turnover and limited follow-up from the project team owing to travel restrictions imposed at the onset of the COVID-19 pandemic reduced knowledge of the systems change among service staff. This contributed to a reduction in the shared accountability required to sustain the components of the systems change that had not already been successfully normalized. The electronic prompts also required moderate digital literacy on the part of individual GPs, and frequent staffing changes and low numbers of long-term ongoing GPs prevented technological upskilling of the GP workforce as a whole:

> *'Cause we have so many doctors coming and going, we do have a couple of local GPs as well. GPs that have been here for years. You know, they're here for 4 weeks, they're away 4 weeks, back for 4 weeks. And then obviously, you know, we have locums which come and go, you know. They're here for 2 weeks. They're gone. And then you get the next one. So, I think it's just basically having so many staff coming and going, and having the availability I guess to, to train them or to go over all that sort of stuff.* [Participant 5, AHP]

Of the 4 components, the presigned pathology forms component was perceived to diverge the most from existing practice, requiring cultural as well as process change. This finding was of interest to us because it did not reflect the outcomes of the initial collaborative development of the systems change. During the planning process, this component had been perceived by service staff as an "easy win," but retrospectively it was seen by participants as a significant and extremely challenging change.

The enhanced data reporting was initially implemented successfully and sustained during the study period but was not sustained during the COVID-19 pandemic. However, participants perceived benefits to reinstating this component in

a post–COVID-19 context. Participants recognized the normalization of STI testing as an ongoing investment in community health and continued to view this as an important priority even in the context of a recent and ongoing crisis that disrupted previously normalized practice.

Reflexive Monitoring

The electronic prompts and automated pathology sets had been proposed by staff as a strategy with high potential to reduce friction and minimize provider-end barriers to increasing comprehensive STI testing. However, although participants felt that the component had realized some of its potential among both existing staff of the service and some newcomers, and the quantitative findings indicate that it had successfully increased STI testing within the health assessment, the capacity for universal normalization was limited by EMR functionality requiring separate configuration for each individual GP user. Participants identified an opportunity to reconfigure the component so that it is implemented from the level of either the software vendor or the pathology level rather than on the practice level and believed that this greater systematization would enable the replicability and sustainability of the observed increase in STI testing:

> *Putting the prompt there, it's easier to one-click for the GPs, to select it, and we will get all the tests that we need, and make sure that it gets reported correctly from the path lab. So we'll add the STI testing, you know, thinking we'll just go for chlamydia, get the urine. But having it on the [autopopulated] panel it says "syphilis," so, therefore, it goes back to either the pathology collector or the nurse, or health worker; they'll immediately draw the blood...I think that's worthwhile.* [Participant 2, manager]

The credentialing component was not integrated into regular practice, explaining our finding of limited increase in overall testing across all consultation types. Participants identified value in introducing a defined clinical educator role to support the training of RNs and the alignment of clinical care with policy. Managers reported that the lack of support from GPs for the presigned pathology forms component dampened the enthusiasm of RNs but believed that the explicit delegation of a GP by management to presign forms would be acceptable to the GP workforce.

Participants perceived a link between the data component and existing and prospective internal communications mechanisms, which they identified as an opportunity to strengthen and sustain the systems change. Recommendations included incorporating data reporting into existing meeting agendas and building on existing channels to streamline communication between management and frontline staff (eg, via a staff newsletter).

Discussion

Principal Findings

This study provides evidence for the first time that optimizing the health assessment is an effective strategy to increase STI testing in the ACCHS setting. It also contributes to the evidence base on the use of electronic prompts and automated pathology sets to increase STI testing, demonstrating that these are effective in the ACCHS setting and acceptable to ACCHS staff. Our qualitative data show that across all 4 components of the systems change, the electronic prompts and automated pathology sets had the highest level of integration and normalization into routine practice, resulting in a substantial and sustained increase in STI testing within the health assessment seen in our quantitative data.

In our evaluation, we observed an increase in overall STI testing integrated into the health assessment in the period when the electronic prompts and automated pathology sets were introduced, from 16% (38/237) in year 1 to 42.9% (94/219) in the intervention period. Although an increase in syphilis testing was observed both within and outside of the health assessment, this was not accompanied by an increase in CT/NG tests outside of the health assessment. As no effect was observed for the internal control, and the intervention coincided with an immediate increase in STI testing, we can conclude that the increases are likely to have been the result of the electronic prompts and automated pathology test sets implemented as part of the systems change. The increase in STI testing generally was primarily seen within health assessment consultations, and the electronic prompt was the only component of the systems change to focus directly on the health assessment, with other components focusing on other consultation types. The increase in syphilis testing outside of the health assessment occurred independently of an increase in tests for other STIs, indicating that this effect represents the inclusion of syphilis tests in existing STI test requests. This suggests that the electronic prompts and automated pathology test sets and their resulting normalization into routine practice were the components of the systems change responsible for the increase in STI testing observed within health assessment consultations and the increase in syphilis testing specifically.

The successful adoption of STI testing into the health assessment by providers and patients is likely to be due to the perceived acceptability and cultural safety of the strategy [57]. Age-based STI screening in routine medical consultations is perceived as acceptable by both GPs and their male and female patients [58-60], which may have contributed to strong support for this strategy among our interview participants. However, the electronic component of the systems change could have achieved greater reach if more health assessments had been conducted among young people. We found that only a quarter of the young people received a health assessment in both the pre- and postintervention periods, consistent with another study in the same context [17]. This meant that despite the increases in STI testing within the health assessment, we only saw a modest increase in STI testing overall, consistent with a previous CQI program involving enhanced data reporting and incentives that increased STI testing among young Aboriginal people within health assessments but not in routine medical consultations outside of the health assessment [61]. Fully realizing the potential of the electronic optimization of the health assessment to increase STI testing will require investment in strategies to increase the uptake of the health assessment itself. This will require investment in both provider-end systems change and better resourcing of health promotion to increase patient demand

[62]. Currently, the health service participates in a health promotion program called Deadly Choices that provides nonfinancial incentives to patients for participating in a health assessment. The role of nonfinancial incentives in increasing health assessment uptake is the focus of a future paper.

Our interview participants valued the work done by champions to push the systems change forward and consistently advocated further investment in dedicated champions to advance project goals. However, the qualitative data also provided evidence of the vulnerabilities created when systems change relies on individuals rather than a sustained, collective, whole-of-service approach. Two components that aimed to increase STI testing outside of the health assessment—the credentialing of RNs and AHPs as well as the presigning of pathology request forms—were not successfully implemented, which explains why STI testing increased substantially within the health assessment but only modestly outside of it. In the primary care setting, RNs tend not to hold positions of equal power to GPs or engage in shared decision-making, and in the Australian context specifically, the inability of RNs to claim reimbursement from Medicare for clinical services has been identified as a significant barrier to increasing shared care [63]. Previous research examining RN-led cervical screening using a model similar to the one discussed in this paper identified medicolegal implications in GPs using their provider number to authorize tests ordered by RNs because there is no way to identify the clinician responsible for requesting an individual test [64]. Our interview participants attributed the unsuccessful normalization of these 2 components primarily to the loss of momentum caused by the departure of key champions. The value of champions in driving forward innovations has been recognized in ACCHS and other primary care settings [65,66], and Aboriginal leadership, in particular, remains an often untapped resource [67]. However, the *National Framework for Continuous Quality Improvement in Primary Health Care for Aboriginal and Torres Strait Islander People 2018-2023* [68] advocates for systems change to be championed at the organizational level rather than by individual staff and recommends that leadership encourage a culture of "CQI is everyone's business."

We also found that a higher proportion of health assessments conducted with men included an STI test in both the pre- and postintervention periods, compared with women, and the increase in the proportion of health assessments that included an STI test in the postintervention period was greater for men than for women. In addition, the proportion of health assessments inclusive of a chlamydia test that had a concurrent syphilis test increased more substantially in the postintervention period for men than for women. Although other studies have instead found that the uptake of STI testing overall is higher in young Aboriginal women [1,17,69-73], our previous work has identified gender parity in STI testing in the context of the health assessment [74]. One potential explanation for this identified in the qualitative data was the impact of standardization resulting in routine tests being offered during the health assessment that clinicians may have previously used their discretion to skip. Qualitative research with Aboriginal men has identified not only a reluctance to proactively access health care [75] but also strongly held values for being a responsible sexual partner [76], highlighting the importance of health assessment in providing routine, comprehensive STI testing to young Aboriginal men who may not otherwise seek out testing.

Normalizing STI testing into routine practice is the cumulative product of multiple small and interrelated changes [77], and our qualitative findings support the value of this model for promoting changes in clinical practice related to STI testing owing to the varied and individual needs of clinicians targeted by systems change [78]. Collaboratively designing this systems change ensured high acceptability even of those components that were not successfully implemented or sustained. Previous literature has found that RN-led testing increases STI testing in primary care, while also noting difficulty in sustaining the required changes to the supporting infrastructure and processes [77]. Although some RNs perceive GPs' reluctance to share care to be a barrier to implementing RN-led testing [39], related research has found that GPs support RN-led testing and believe that it is suitable for their skill set [79]. Our findings suggest potential for more successful future implementation of these components of the systems change, with appropriate investment in ongoing engagement of staff at all levels. Future work using NPT may benefit from examining the elements of collaboration associated with successful systems change and identifying what may satisfy expectations to achieve normalization.

Our evaluation has a few limitations to consider when interpreting the findings. First, the qualitative data were collected over a longer period than was planned after the conclusion of the postintervention period owing to the impacts of the COVID-19 pandemic on the participating ACCHS. Although it is also a standard expectation that qualitative findings be recognized as nongeneralizable owing to the small samples typical in qualitative research, the qualitative sample size represented more than half of the staff who had continued employment throughout the study period, which supports the value of incorporating these findings into the analysis as a whole. Second, our quantitative analysis used a before-and-after design to evaluate the effectiveness of the intervention, which may be subject to temporal biases but was strengthened using an ITS analysis with a 2-year before period (12 time points) and an internal control. Third, we were unable to evaluate which GPs installed the prompts and automated pathology sets or how frequently they were activated. Finally, the strategy was evaluated at 1 health service; therefore, the findings may not be generalizable to all NSW ACCHSs. The use of the NPT framework is a strength of this study because a systematic review has identified that NPT strengthens analysis of implementation processes [28].

Conclusions

We have shown that the introduction to the EMR of an electronic prompt to encourage GPs to offer age-based STI testing and a shortcut to autopopulate a pathology request for a comprehensive STI test is an effective and acceptable strategy to increase both integration of STI tests into the health assessment and the inclusion of syphilis testing in STI tests completed within the health assessment in the ACCHS setting. Future systems change should also incorporate strategies to increase the uptake of the health assessment among the target

population and work to better promote whole-of-service engagement and accountability to increase normalization. Future work exploring the suitability of related interventions for other sites should continue collaboration with ACCHS staff to develop locally customized solutions.

Conflicts of Interest

None declared.

References

1. Ward J, Bryant J, Worth H, Hull P, Solar S, Bailey S. Use of health services for sexually transmitted and blood-borne viral infections by young Aboriginal people in New South Wales. Aust J Prim Health 2013;19(1):81-86. [doi: 10.1071/PY11032] [Medline: 22951105]
2. Harfield SG, Davy C, McArthur A, Munn Z, Brown A, Brown N. Characteristics of indigenous primary health care service delivery models: a systematic scoping review. Global Health 2018 Jan 25;14(1):12 [FREE Full text] [doi: 10.1186/s12992-018-0332-2] [Medline: 29368657]
3. Hutchinson P, Tobin P, Muirhead A, Robinson N. Closing the gaps in cancer screening with first nations, Inuit, and Métis populations: a narrative literature review. J Indig Wellbeing 2018;3(1):3-17 [FREE Full text]
4. Palmer SC, Gray H, Huria T, Lacey C, Beckert L, Pitama SG. Reported Māori consumer experiences of health systems and programs in qualitative research: a systematic review with meta-synthesis. Int J Equity Health 2019 Oct 28;18(1):163 [FREE Full text] [doi: 10.1186/s12939-019-1057-4] [Medline: 31660988]
5. Allen L, Hatala A, Ijaz S, Courchene ED, Bushie EB. Indigenous-led health care partnerships in Canada. CMAJ 2020 Mar 02;192(9):E208-E216. [doi: 10.1503/cmaj.190728] [Medline: 32122977]
6. Davy C, Harfield S, McArthur A, Munn Z, Brown A. Access to primary health care services for Indigenous peoples: a framework synthesis. Int J Equity Health 2016 Sep 30;15(1):163 [FREE Full text] [doi: 10.1186/s12939-016-0450-5] [Medline: 27716235]
7. Guy R, Ward JS, Smith KS, Su JY, Huang RL, Tangey A, et al. The impact of sexually transmissible infection programs in remote aboriginal communities in Australia: a systematic review. Sex Health 2012 Jul;9(3):205-212. [doi: 10.1071/SH11074] [Medline: 22697136]
8. Bloodborne viral and sexually transmissible infections in Aboriginal and Torres Strait Islander people: annual surveillance report 2018. Kirby Institute, UNSW Sydney. 2018. URL: https://www.kirby.unsw.edu.au/sites/default/files/documents/KI_Aboriginal-Surveillance-Report-2018.pdf [accessed 2023-11-19]
9. NSW sexually transmissible infections strategy January to December 2020 data report. NSW Ministry of Health. 2021. URL: https://www.health.nsw.gov.au/Infectious/Reports/Publications/sti/nsw-2020-sti-report.pdf [accessed 2023-11-19]
10. Korenromp EL, Sudaryo MK, de Vlas SJ, Gray RH, Sewankambo NK, Serwadda D, et al. What proportion of episodes of gonorrhoea and chlamydia becomes symptomatic? Int J STD AIDS 2002 Feb;13(2):91-101. [doi: 10.1258/0956462021924712] [Medline: 11839163]
11. Wijesooriya NS, Rochat RW, Kamb ML, Turlapati P, Temmerman M, Broutet N, et al. Global burden of maternal and congenital syphilis in 2008 and 2012: a health systems modelling study. Lancet Glob Health 2016 Aug;4(8):e525-e533 [FREE Full text] [doi: 10.1016/S2214-109X(16)30135-8] [Medline: 27443780]
12. Reekie J, Donovan B, Guy R, Hocking JS, Jorm L, Kaldor JM, et al. Hospitalisations for pelvic inflammatory disease temporally related to a diagnosis of Chlamydia or gonorrhoea: a retrospective cohort study. PLoS One 2014 Apr 17;9(4):e94361 [FREE Full text] [doi: 10.1371/journal.pone.0094361] [Medline: 24743388]
13. Reekie J, Donovan B, Guy R, Hocking JS, Kaldor JM, Mak D, Chlamydia and Reproductive Health Outcome Investigators. Risk of ectopic pregnancy and tubal infertility following gonorrhea and chlamydia infections. Clin Infect Dis 2019 Oct 15;69(9):1621-1623. [doi: 10.1093/cid/ciz145] [Medline: 30778532]
14. Hui BB, Wilson DP, Ward JS, Guy RJ, Kaldor JM, Law MG, et al. The potential impact of new generation molecular point-of-care tests on gonorrhoea and chlamydia in a setting of high endemic prevalence. Sex Health 2013 Aug;10(4):348-356. [doi: 10.1071/SH13026] [Medline: 23806149]
15. Thomas SL, Kalsi H, Graham S. Fifth National Aboriginal and Torres Strait Islander blood-borne virus and sexually transmissible infections strategy 2018-2022. Australian Government Department of Health. 2012. URL: https://www.health.gov.au/sites/default/files/documents/2022/06/fifth-national-aboriginal-and-torres-strait-islander-bloodborne-viruses-and-sexually-transmissible-infections-strategy-2018-2022.pdf [accessed 2023-11-19]
16. NSW sexually transmissible infections strategy 2022-2026. NSW Ministry of Health. 2022. URL: https://www.health.nsw.gov.au/sexualhealth/Publications/nsw-sti-strategy.pdf [accessed 2023-11-19]
17. Nattabi B, Matthews V, Bailie J, Rumbold A, Scrimgeour D, Schierhout G, et al. Wide variation in sexually transmitted infection testing and counselling at Aboriginal primary health care centres in Australia: analysis of longitudinal continuous quality improvement data. BMC Infect Dis 2017 Feb 15;17(1):148 [FREE Full text] [doi: 10.1186/s12879-017-2241-z] [Medline: 28201979]

18. Practice incentives program: Indigenous health incentive. Australian Government Department of Health. URL: https://www.health.gov.au/our-work/practice-incentives-program-indigenous-health-incentive [accessed 2023-11-18]
19. National guide to a preventive health assessment for Aboriginal and Torres Strait Islander people. National Aboriginal Community Controlled Health Organisation and The Royal Australian College of General Practitioners. 2018. URL: https://www.racgp.org.au/FSDEDEV/media/documents/Clinical%20Resources/Resources/National-guide-3rd-ed-Sept-2018-web.pdf [accessed 2023-11-18]
20. Health checks and follow-ups for Aboriginal and Torres Strait Islander people. Australian Institute of Health and Welfare. 2022. URL: https://www.aihw.gov.au/reports/indigenous-australians/indigenous-health-checks-follow-ups/contents/rate-of-health-checks/differences-by-sex-and-age [accessed 2023-11-18]
21. Newman CE, Fraser D, Ong JJ, Bourne C, Grulich AE, Bavinton BR. Sustaining sexual and reproductive health through COVID-19 pandemic restrictions: qualitative interviews with Australian clinicians. Sex Health 2022 Dec;19(6):525-532. [doi: 10.1071/SH22109] [Medline: 36038359]
22. Executive summary: Aboriginal and Torres Strait Islander health checks: results from testing in health services and general practices. National Aboriginal Community Controlled Health Organisation and The Royal Australian College of General Practitioners. 2021. URL: https://www.racgp.org.au/FSDEDEV/media/documents/Faculties/ATSI/Exec-Summary-NACCHO-RACGP-Health-checks-testing-report-2021.pdf [accessed 2023-11-18]
23. Spurling GK, Askew DA, Schluter PJ, Hayman NE. Implementing computerised Aboriginal and Torres Strait Islander health checks in primary care for clinical care and research: a process evaluation. BMC Med Inform Decis Mak 2013 Sep 21;13:108 [FREE Full text] [doi: 10.1186/1472-6947-13-108] [Medline: 24053425]
24. Guy RJ, Ali H, Liu B, Poznanski S, Ward J, Donovan B, et al. Efficacy of interventions to increase the uptake of chlamydia screening in primary care: a systematic review. BMC Infect Dis 2011 Aug 05;11:211 [FREE Full text] [doi: 10.1186/1471-2334-11-211] [Medline: 21816113]
25. Callander D, Bourne C, Wand H, Stoové M, Hocking JS, de Wit J, et al. Assessing the impacts of integrated decision support software on sexual orientation recording, comprehensive sexual health testing, and detection of infections among gay and bisexual men attending general practice: observational study. JMIR Med Inform 2018 Nov 06;6(4):e10808 [FREE Full text] [doi: 10.2196/10808] [Medline: 30401672]
26. Bissessor M, Fairley CK, Leslie D, Chen MY. Use of a computer alert increases detection of early, asymptomatic syphilis among higher-risk men who have sex with men. Clin Infect Dis 2011 Jul 01;53(1):57-58. [doi: 10.1093/cid/cir271] [Medline: 21653303]
27. Guy R, El-Hayek C, Fairley CK, Wand H, Carr A, McNulty A, et al. Opt-out and opt-in testing increases syphilis screening of HIV-positive men who have sex with men in Australia. PLoS One 2013;8(8):e71436 [FREE Full text] [doi: 10.1371/journal.pone.0071436] [Medline: 24009661]
28. May CR, Cummings A, Girling M, Bracher M, Mair FS, May CM, et al. Using normalization process theory in feasibility studies and process evaluations of complex healthcare interventions: a systematic review. Implement Sci 2018 Jun 07;13(1):80 [FREE Full text] [doi: 10.1186/s13012-018-0758-1] [Medline: 29879986]
29. May CR, Mair F, Finch T, MacFarlane A, Dowrick C, Treweek S, et al. Development of a theory of implementation and integration: normalization process theory. Implement Sci 2009 May 21;4:29 [FREE Full text] [doi: 10.1186/1748-5908-4-29] [Medline: 19460163]
30. May C. Towards a general theory of implementation. Implement Sci 2013 Feb 13;8:18 [FREE Full text] [doi: 10.1186/1748-5908-8-18] [Medline: 23406398]
31. Murray E, Treweek S, Pope C, MacFarlane A, Ballini L, Dowrick C, et al. Normalisation process theory: a framework for developing, evaluating and implementing complex interventions. BMC Med 2010 Oct 20;8:63 [FREE Full text] [doi: 10.1186/1741-7015-8-63] [Medline: 20961442]
32. Mid-term evaluation of the NSW Aboriginal health plan 2013-2023. NSW Ministry Of Health. 2019. URL: https://www.health.nsw.gov.au/research/Publications/ahp-mid-term-main-report.pdf [accessed 2023-11-18]
33. Broken hill: 2016 census all persons QuickStats. Australian Bureau of Statistics. 2016. URL: https://www.abs.gov.au/census/find-census-data/quickstats/2016/SSC10592 [accessed 2023-11-18]
34. Wilcannia: 2016 Census All persons QuickStats. Australian Bureau of Statistics. 2016. URL: https://www.abs.gov.au/census/find-census-data/quickstats/2016/SSC14290
35. Azzopardi PS, Kennedy EC, Patton GC, Power R, Roseby RD, Sawyer SM, et al. The quality of health research for young Indigenous Australians: systematic review. Med J Aust 2013 Jul 08;199(1):57-63. [doi: 10.5694/mja12.11141] [Medline: 23829266]
36. Graham S, Martin K, Gardner K, Beadman M, Doyle MF, Bolt R, et al. Aboriginal young people's perspectives and experiences of accessing sexual health services and sex education in Australia: a qualitative study. Glob Public Health 2023 Jan;18(1):2196561 [FREE Full text] [doi: 10.1080/17441692.2023.2196561] [Medline: 37018760]
37. Spurway K, Sullivan C, Leha J, Trewlynn W, Briskman L, Soldatic K. “I felt invisible”: first nations LGBTIQSB+ young people’s experiences with health service provision in Australia. J Gay Lesbian Soc Serv 2022 Mar 08;35(1):68-91 [FREE Full text] [doi: 10.1080/10538720.2022.2045241]

38. STI testing and management advice has changed. Australasian Society for HIV, Viral Hepatitis and Sexual Health Medicine. 2022. URL: https://sti.guidelines.org.au/ [accessed 2023-11-18]
39. Lorch R, Hocking J, Guy R, Vaisey A, Wood A, Lewis D, ACCEPt Consortium. Practice nurse chlamydia testing in Australian general practice: a qualitative study of benefits, barriers and facilitators. BMC Fam Pract 2015 Mar 14;16:36 [FREE Full text] [doi: 10.1186/s12875-015-0251-8] [Medline: 25880077]
40. Gunaratnam P, Schierhout G, Brands J, Maher L, Bailie R, Ward J, et al. Qualitative perspectives on the sustainability of sexual health continuous quality improvement in clinics serving remote Aboriginal communities in Australia. BMJ Open 2019 May 05;9(5):e026679 [FREE Full text] [doi: 10.1136/bmjopen-2018-026679] [Medline: 31061040]
41. Hocking JS, Wood A, Temple-Smith M, Braat S, Law M, Bulfone L, et al. The impact of removing financial incentives and/or audit and feedback on chlamydia testing in general practice: a cluster randomised controlled trial (ACCEPt-able). PLoS Med 2022 Jan;19(1):e1003858 [FREE Full text] [doi: 10.1371/journal.pmed.1003858] [Medline: 34982767]
42. McCormack H, Guy R, Bourne C, Newman CE. Integrating testing for sexually transmissible infections into routine primary care for Aboriginal young people: a strengths-based qualitative analysis. Aust N Z J Public Health 2022 Jun;46(3):370-376 [FREE Full text] [doi: 10.1111/1753-6405.13208] [Medline: 35238454]
43. Fogarty W, Lovell M, Langenberg J, Heron MJ. Deficit discourse and strengths-based approaches: changing the narrative of Aboriginal and Torres Strait Islander health and wellbeing. The Lowitja Institute. URL: https://www.lowitja.org.au/content/Document/Lowitja-Publishing/deficit-discourse-strengths-based.pdf [accessed 2023-11-18]
44. Thurber KA, Thandrayen J, Banks E, Doery K, Sedgwick M, Lovett R. Strengths-based approaches for quantitative data analysis: a case study using the Australian Longitudinal Study of Indigenous Children. SSM Popul Health 2020 Dec;12:100637 [FREE Full text] [doi: 10.1016/j.ssmph.2020.100637] [Medline: 32923575]
45. Creswell JW, Clark VL. Designing and Conducting Mixed Methods Research. Thousand Oaks, CA: Sage Publications; 2017.
46. Boyle D, Kong F. A systematic mechanism for the collection and interpretation of display format pathology test results from Australian primary care records. Electron J Health Inform 2011;6(2):e18.
47. Bradley C, Hengel B, Crawford K, Elliott S, Donovan B, Mak DB, et al. Establishment of a sentinel surveillance network for sexually transmissible infections and blood borne viruses in Aboriginal primary care services across Australia: the ATLAS project. BMC Health Serv Res 2020 Aug 20;20(1):769 [FREE Full text] [doi: 10.1186/s12913-020-05388-y] [Medline: 32819360]
48. Kontopantelis E, Doran T, Springate DA, Buchan I, Reeves D. Regression based quasi-experimental approach when randomisation is not an option: interrupted time series analysis. BMJ 2015 Jun 09;350:h2750 [FREE Full text] [doi: 10.1136/bmj.h2750] [Medline: 26058820]
49. Guest G, Namey EE, Mitchell M. Collecting Qualitative Data: A Field Manual for Applied Research. Thousand Oaks, CA: Sage Publications; 2013.
50. May C, Finch T. Implementing, embedding, and integrating practices: an outline of normalization process theory. Sociol 2009 Jun 15;43(3):535-554 [FREE Full text] [doi: 10.1177/0038038509103208]
51. Leon N, Lewin S, Mathews C. Implementing a provider-initiated testing and counselling (PITC) intervention in Cape town, South Africa: a process evaluation using the normalisation process model. Implement Sci 2013 Aug 26;8:97 [FREE Full text] [doi: 10.1186/1748-5908-8-97] [Medline: 23972055]
52. Hengel B, Wand H, Ward J, Rumbold A, Garton L, Taylor-Thomson D, STRIVE Investigators. Patient, staffing and health centre factors associated with annual testing for sexually transmissible infections in remote primary health centres. Sex Health 2017 Jun;14(3):274-281. [doi: 10.1071/SH16123] [Medline: 28445684]
53. Hengel B, Bell S, Garton L, Ward J, Rumbold A, Taylor-Thomson D, STRIVE Investigators. Perspectives of primary health care staff on the implementation of a sexual health quality improvement program: a qualitative study in remote aboriginal communities in Australia. BMC Health Serv Res 2018 Apr 02;18(1):230 [FREE Full text] [doi: 10.1186/s12913-018-3024-y] [Medline: 29609656]
54. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
55. Braun V, Clarke V. Thematic analysis. In: Cooper H, Camic PM, Long DL, Panter AT, Rindskopf D, Sher KJ, editors. APA Handbook of Research Methods in Psychology, Vol. 2. Research Designs: Quantitative, Qualitative, Neuropsychological, and Biological. New York, NY: American Psychological Association; 2012:57-71.
56. Braun V, Clarke V. Successful Qualitative Research: A Practical Guide for Beginners. Thousand Oaks, CA: Sage Publications; 2013.
57. Hickey S, Roe Y, Harvey C, Kruske S, Clifford-Motopi A, Fisher I, et al. Community-based sexual and reproductive health promotion and services for first nations people in urban Australia. Int J Womens Health 2021;13:467-478 [FREE Full text] [doi: 10.2147/IJWH.S297479] [Medline: 34040455]
58. Pavlin NL, Parker R, Fairley CK, Gunn JM, Hocking J. Take the sex out of STI screening! Views of young women on implementing chlamydia screening in General Practice. BMC Infect Dis 2008 May 09;8:62 [FREE Full text] [doi: 10.1186/1471-2334-8-62] [Medline: 18471280]

59. Latreille S, Collyer A, Temple-Smith M. Finding a segue into sex: young men's views on discussing sexual health with a GP. Aust Fam Physician 2014 Apr;43(4):217-221 [FREE Full text] [Medline: 24701626]
60. Hocking JS, Parker RM, Pavlin N, Fairley CK, Gunn JM. What needs to change to increase chlamydia screening in general practice in Australia? The views of general practitioners. BMC Public Health 2008 Dec 30;8:425 [FREE Full text] [doi: 10.1186/1471-2458-8-425] [Medline: 19115998]
61. Graham S, Guy RJ, Wand HC, Kaldor JM, Donovan B, Knox J, et al. A sexual health quality improvement program (SHIMMER) triples chlamydia and gonorrhoea testing rates among young people attending Aboriginal primary health care services in Australia. BMC Infect Dis 2015 Sep 02;15:370 [FREE Full text] [doi: 10.1186/s12879-015-1107-5] [Medline: 26329123]
62. Jennings W, Spurling GK, Askew DA. Yarning about health checks: barriers and enablers in an urban Aboriginal medical service. Aust J Prim Health 2014;20(2):151-157. [doi: 10.1071/PY12138] [Medline: 23552601]
63. McInnes S, Peters K, Bonney A, Halcomb E. An integrative review of facilitators and barriers influencing collaboration and teamwork between general practitioners and nurses working in general practice. J Adv Nurs 2015 Sep;71(9):1973-1985. [doi: 10.1111/jan.12647] [Medline: 25731727]
64. Mills J, Chamberlain-Salaun J, Christie L, Kingston M, Gorman E, Harvey C. Australian nurses in general practice, enabling the provision of cervical screening and well women's health care services: a qualitative study. BMC Nurs 2012 Nov 12;11:23 [FREE Full text] [doi: 10.1186/1472-6955-11-23] [Medline: 23145901]
65. O'Connell E, Hogan A, Ricketts E, Jacomelli J, McNulty C. Advantages of chlamydia screening in general practice settings. Prim Health Care 2013 May;23(4):26-29. [doi: 10.7748/phc2013.05.23.4.26.e710]
66. MacDonald C, Genat B, Thorpe S, Browne J. Establishing health-promoting workplaces in Aboriginal community organisations: healthy eating policies. Aust J Prim Health 2016;22(3):239-243. [doi: 10.1071/PY14144] [Medline: 25720592]
67. Stroud V, Adams J, Champion D, Hogarth G, Mahony A, Monck R, et al. The role of Aboriginal leadership in community health programmes. Prim Health Care Res Dev 2021 Oct 29;22:E58 [FREE Full text] [doi: 10.1017/s1463423621000414]
68. National framework for continuous quality improvement in primary health care for Aboriginal and Torres Strait Islander people 2018-2023. National Aboriginal Community Controlled Health Organisation. URL: https://www.naccho.org.au/app/uploads/2022/03/NACCHO-CQI-Framework-2019-1.pdf [accessed 2023-11-18]
69. Bell S, Aggleton P, Ward J, Murray W, Silver B, Lockyer A, et al. Young Aboriginal people's engagement with STI testing in the Northern Territory, Australia. BMC Public Health 2020 Apr 06;20(1):459 [FREE Full text] [doi: 10.1186/s12889-020-08565-0] [Medline: 32252712]
70. Graham S, Smith LW, Fairley CK, Hocking J. Prevalence of chlamydia, gonorrhoea, syphilis and trichomonas in Aboriginal and Torres Strait Islander Australians: a systematic review and meta-analysis. Sex Health 2016 Apr;13(2):99-113. [doi: 10.1071/SH15171] [Medline: 26775118]
71. Harrod ME, Couzos S, Ward J, Saunders M, Donovan B, Hammond B, et al. Gonorrhoea testing and positivity in non-remote Aboriginal Community Controlled Health Services. Sex Health 2017 Aug;14(4):320-324. [doi: 10.1071/SH16046] [Medline: 28641073]
72. Graham S, Wand HC, Ward JS, Knox J, McCowen D, Bullen P, et al. Attendance patterns and chlamydia and gonorrhoea testing among young people in Aboriginal primary health centres in New South Wales, Australia. Sex Health 2015 Oct;12(5):445-452. [doi: 10.1071/SH15007] [Medline: 26210444]
73. Ward J, Goller J, Ali H, Bowring A, Couzos S, Saunders M, ACCESS Collaboration. Chlamydia among Australian Aboriginal and/or Torres Strait Islander people attending sexual health services, general practices and Aboriginal Community Controlled Health Services. BMC Health Serv Res 2014 Jul 01;14:285 [FREE Full text] [doi: 10.1186/1472-6963-14-285] [Medline: 24981418]
74. McCormack H, Wand H, Bourne C, Ward J, Bradley C, Mak D, et al. Integrating testing for sexually transmissible infections into annual health assessments for Aboriginal and Torres Strait Islander young people: a cross-sectional analysis. Sex Health 2023 Sep 11 [FREE Full text] [doi: 10.1071/SH23107] [Medline: 37690512]
75. Newman CE, Gray R, Brener L, Jackson LC, Dillon A, Saunders V, et al. "I had a little bit of a bloke meltdown…but the next day, I was up": understanding cancer experiences among aboriginal men. Cancer Nurs 2017 May;40(3):E1-E8. [doi: 10.1097/NCC.0000000000000399] [Medline: 27271367]
76. Graham S, Martin K, Beadman M, Doyle M, Bolt R. Our relationships, our values, our culture - Aboriginal young men's perspectives about sex, relationships and gender stereotypes in Australia. Cult Health Sex 2023 Mar;25(3):304-319. [doi: 10.1080/13691058.2022.2039776] [Medline: 35192437]
77. Yeung A, Temple-Smith M, Fairley C, Hocking J. Narrative review of the barriers and facilitators to chlamydia testing in general practice. Aust J Prim Health 2015;21(2):139-147. [doi: 10.1071/PY13158] [Medline: 25118823]
78. Yeung A, Hocking J, Guy R, Fairley CK, Smith K, Vaisey A, ACCEPt consortium. 'It Opened My Eyes'-examining the impact of a multifaceted chlamydia testing intervention on general practitioners using Normalization Process Theory. Fam Pract 2018 Sep 18;35(5):626-632. [doi: 10.1093/fampra/cmy011] [Medline: 29608672]
79. Lorch R, Hocking J, Guy R, Vaisey A, Wood A, Donovan B, ACCEPt consortium. Do Australian general practitioners believe practice nurses can take a role in chlamydia testing? A qualitative study of attitudes and opinions. BMC Infect Dis 2015 Jan 31;15:31 [FREE Full text] [doi: 10.1186/s12879-015-0757-7] [Medline: 25885341]

Abbreviations

ACCHS: Aboriginal Community Controlled Health Service
AHP: Aboriginal health practitioner
CQI: continuous quality improvement
CT: *Chlamydia trachomatis*
EMR: electronic medical record
GP: general practitioner
GRHANITE: Generic Health Network Information Technology for the Enterprise
ITS: interrupted time series
NG: *Neisseria gonorrhea*
NPT: normalization process theory
NSW: New South Wales
RN: registered nurse
STI: sexually transmitted infection

Edited by J Hefner; submitted 30.07.23; peer-reviewed by J Walsh, O Petrovskaya, J Hou; comments to author 13.09.23; revised version received 22.10.23; accepted 13.11.23; published 30.11.23.

Please cite as:
McCormack H, Wand H, Newman CE, Bourne C, Kennedy C, Guy R
Exploring Whether the Electronic Optimization of Routine Health Assessments Can Increase Testing for Sexually Transmitted Infections and Provider Acceptability at an Aboriginal Community Controlled Health Service: Mixed Methods Evaluation
JMIR Med Inform 2023;11:e51387
URL: https://medinform.jmir.org/2023/1/e51387
doi:10.2196/51387
PMID:38032729

How People With a Bipolar Disorder Diagnosis Talk About Personal Recovery in Peer Online Support Forums: Corpus Framework Analysis Using the POETIC Framework

Glorianna Jagfeld[1,2], PhD; Fiona Lobban[1], PhD; Chloe Humphreys[3], BA; Paul Rayson[2], PhD; Steven Huntley Jones[1], PhD

1

2

3

Corresponding Author:
Steven Huntley Jones, PhD

Abstract

Background: Personal recovery is of particular value in bipolar disorder, where symptoms often persist despite treatment. We previously defined the POETIC (Purpose and Meaning, Optimism and Hope, Empowerment, Tensions, Identity, Connectedness) framework for personal recovery in bipolar disorder. So far, personal recovery has only been studied in researcher-constructed environments (eg, interviews and focus groups). Support forum posts can serve as a complementary naturalistic data resource to understand the lived experience of personal recovery.

Objective: This study aimed to answer the question "What can online support forum posts reveal about the experience of personal recovery in bipolar disorder in relation to the POETIC framework?"

Methods: By integrating natural language processing, corpus linguistics, and health research methods, this study analyzed public, bipolar disorder support forum posts relevant to the lived experience of personal recovery. By comparing 4462 personal recovery–relevant posts by 1982 users to 25,197 posts not relevant to personal recovery, we identified 130 significantly overused key lemmas. Key lemmas were coded according to the POETIC framework.

Results: Personal recovery–related discussions primarily focused on 3 domains: "Purpose and meaning" (particularly reproductive decisions and work), "Connectedness" (romantic relationships and social support), and "Empowerment" (self-management and personal responsibility). This study confirmed the validity of the POETIC framework to capture personal recovery experiences shared on the web and highlighted new aspects beyond previous studies using interviews and focus groups.

Conclusions: This study is the first to analyze naturalistic data on personal recovery in bipolar disorder. By indicating the key areas that people focus on in personal recovery when posting freely and the language they use, this study provides helpful starting points for formal and informal carers to understand the concerns of people diagnosed with a bipolar disorder and to consider how to best offer support.

(JMIR Med Inform 2023;11:e46544) doi:10.2196/46544

KEYWORDS

bipolar disorder; personal recovery; peer online support forums; natural language processing; corpus linguistics; social media; online support; recovery

Introduction

Bipolar disorder (BD) is a severe mental health (MH) problem characterized by recurring episodes of depressed and elevated mood [1]. Its lifetime prevalence ranges from 0.1% to 2.6% internationally [2]. BD is associated with lower quality of life [3] and high suicide risk [4]. Therefore, fostering recovery and living well with BD are important societal tasks.

MH care agendas increasingly focus on enhancing personal recovery (PR), defined as "a way of living a satisfying, hopeful life even with the limitations caused by the illness" [5]. This contrasts with a previously narrower focus on reducing symptoms (clinical recovery). PR might be of particular value in BD [6], where symptoms often persist despite treatment, but has been underresearched to date [7]. Jagfeld et al [8] (hereafter the POETIC review) recently synthesized 12 qualitative studies to develop the first conceptual framework for PR in BD. The POETIC (Purpose and Meaning, Optimism and Hope, Empowerment, Tensions, Identity, Connectedness) framework, based on the CHIME (Connectedness, Hope and Optimism, Identity, Meaning and Purpose, Empowerment) framework [9], comprises the following processes: "Purpose and meaning,"

"Optimism and hope," "Empowerment," "Tensions," "Identity," and "Connectedness" (see Table 1).

Table . The POETIC[a] framework [8]: lived experience of personal recovery in bipolar disorder.

First-level domains	Second-level categories
Purpose and meaning	• Meaning of mental illness experiences • Paid or voluntary work • Quality of life • Meaningful life and social roles
Optimism and hope	• Belief in possibility of recovery • Positive thinking and valuing success • Hope-inspiring relationships • Having dreams and aspirations
Empowerment	• Self-management and personal responsibility • Controversial role of medication • Control over life
Tensions	• Balancing acceptance with ambitions • Openness enables support but also stigmatization • Ambivalence around (hypo)mania
Identity	• Rebuilding positive sense of self • Overcoming stigma • Dimensions of identity
Connectedness	• Support from others • Relationships • Peer support and support groups • Being part of the community

[a]POETIC: Purpose and Meaning, Optimism and Hope, Empowerment, Tensions, Identify, Connectedness.

Current research on PR in BD has several limitations. First, it is mainly based on qualitative studies with few participants [10] and expert opinions, lacking quantitative evidence from larger samples [11]. Second, data collection is limited to structured settings (semistructured interviews, focus groups, and structured measures), which are not naturalistic and are subject to either interviewer bias [12] or constrained responses in structured measures. Third, recruitment is biased toward people who want to talk about PR and are in contact with services or researchers [8].

Naturalistic data collection, where "participants are not aware that they are being studied" [13], overcomes many of these limitations. Online forum posts are a source of naturalistic data, which can offer potential insights into "an experience as it is lived rather than as it is enacted in the researcher constructed environment" [14]. Some natural language processing (NLP) studies have analyzed large numbers of BD online forum posts via automatic quantitative methods such as content analysis [15] or emotion analysis [16,17] to identify forum topics or language differences between people with different or no MH diagnoses. Qualitative studies have applied conversation analysis [18], thematic analysis [19], grounded theory [20], and content analysis [21] to BD online forum posts. Such studies offer rich nuanced accounts of web-based discussions on BD but include only few, often handpicked, posts.

Corpus linguistics [22] provides a mix of quantitative and qualitative methods informed by linguistic theory for analyzing large amounts of text data with depth and richness that can overcome some of the shortcomings of previous NLP and qualitative studies. Semino et al [23] analyzed interviews and online forum posts of patients with cancer and their carers to learn about their lived experience and the metaphors they use for dealing with cancer. Hunt and Brookes [24] applied a combination of corpus linguistics and discourse analysis [25] to MH forum posts. Two corpus-linguistics studies have focused on BD specifically: Abdo et al [26] studied linguistic types of judgments, and McDonald and Woodward-Kron [27] studied forum users' roles and identities.

A systematic review strongly recommended considering web-based content from individuals with lived experience in PR research [10], which has not yet been done. Therefore, the main aim of this paper was to gain further insights into the experience of PR in BD from online forum posts via a combination of NLP, corpus linguistics, and qualitative health research methods. Furthermore, the POETIC framework, synthesized from data collected via interviews or focus groups, has not been applied to new data yet. Hence, the secondary aim of this paper was to validate the framework by exploring to what extent it captures experiences shared on the web. The research question covering both aims is "What can online

support forum posts reveal about the experience of PR in BD in relation to the POETIC framework?"

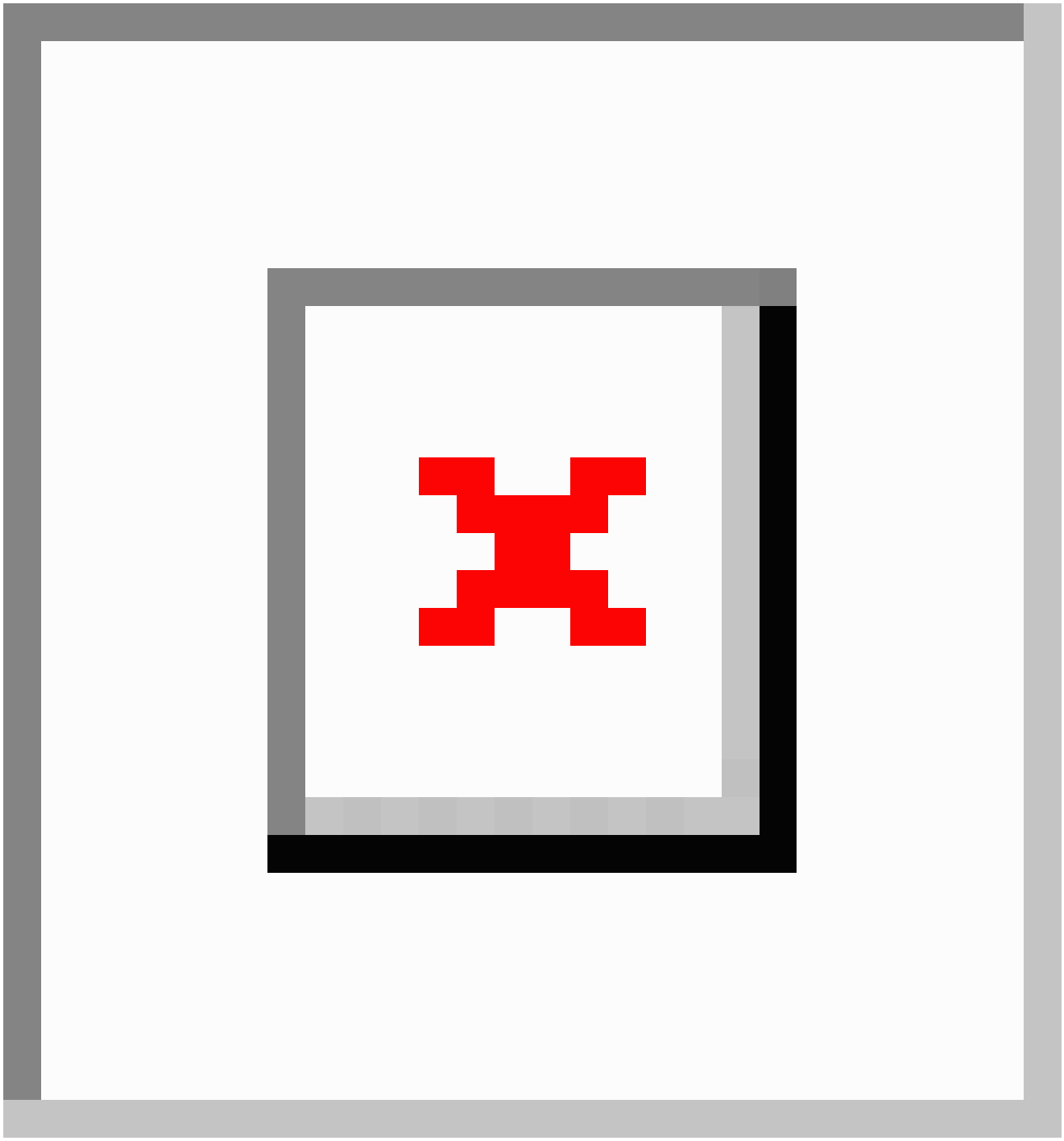
Methods

Data Source

This study analyzed posts from the international web-based discussion platform Reddit [28], which hosts subforums (subreddits) for various topics, including BD. Several reasons motivated the choice of this site: Reddit is one of the most visited internet sites worldwide with an international user base [29]; in contrast to other online support communities, everyone can read all public posts without a user account; and Reddit allows data analysis by third parties.

Reddit users with a self-reported BD diagnosis (S-BiDD) were automatically identified by matching phrases such as "I was diagnosed with bipolar" in all posts between January 2005 (the inception of Reddit) and March 2019 (see Jagfeld et al [30]). All posts of the identified users form the S-BiDD data set. Naturalistic data collection required subsequent filtering for content relevant for PR in BD, as an exploratory study revealed that the posts in the S-BiDD data set covered many other topics (see Report S1 in Multimedia Appendix 1). Figure 1 displays a flow chart for construction of the PR-BD corpus. In linguistics, a corpus is a sampled collection of texts representing a particular language variety [31]. The basis for the corpus was only posts in BD subreddits [32] (fourth level="bipolar"), because a second exploratory study found that references to "recovery" and associated word forms were almost exclusively in relation to BD in BD subreddits (see Report S2 in Multimedia Appendix 1). Furthermore, only posts mentioning BD [33] were selected because only two-thirds (66%) of MH-related "recovery" mentions in BD subreddits referred to BD.

Figure 1. Flowchart of the 4 steps to create the PR-BD corpus and reference corpus. BD: bipolar disorder; MH: mental health; PR: personal recovery; S-BiDD: self-reported bipolar disorder diagnosis.



To select PR-relevant posts, a list of PR terms (comprising both single words and multiword phrases; n=562) [34] was compiled using corpus-linguistics methods (Document S2 in Multimedia Appendix 1). BD subreddit posts that mentioned BD were ranked according to their similarity with the PR terms list via term frequency–inverse document frequency–weighted cosine similarity, a standard information retrieval approach [35,36] (see Document S3 in Multimedia Appendix 1). To determine the cosine similarity cutoff, GJ coded whether 90 posts pertained to PR in BD using a preliminary codebook based on the second exploratory study and the POETIC codebook. SHJ audited the coding. To select the 90 posts, 10 posts were randomly sampled from every 10% quantile of the cosine similarity scores, taking only 10 posts from the first 2 quantiles that all scored 0. Following this, a minimum length of 94 words was set, as 5 posts shorter than this length lacked context to decide on their PR relevance (see Figure S5 in Multimedia Appendix 1). The codebook was refined to its final version (Document S4 in Multimedia Appendix 1). GJ and CH blindly coded 120 additional posts, again randomly sampled from each quantile of the cosine scores.

Ethical Considerations

The Lancaster University Faculty of Health and Medicine research ethics committee approved this research in May 2019 (reference FHMREC18066), which follows ethics guidelines for internet-mediated research [37]. It was infeasible to seek

individual informed consent from the large number of included forum users, but quotes were paraphrased to protect users' anonymity (see Document S5 in Multimedia Appendix 1). We recognize that some people may object to the use of web-based posts as research data without individual consent (eg, [38]). Users generally post to share information or seek support and do not directly provide their content for research. However, we believe that on balance, the benefits of this research to better understand PR makes it worthwhile while acknowledging these potential objections.

The authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008.

Involvement of People With Lived Experience

In all, 4 volunteers with lived experience of BD, who use online forums, and were recruited via People in Research [39] provided input on the study design, results, and subsequent plans in individual web-based meetings. All 4 volunteers—1 man and 3 women—were UK-based and in their 30s to 40s; additionally, 3 (75%) reported a bipolar II disorder diagnosis, 1 (25%) did not further specify their BD diagnosis, and at least 2 (50%) had a migrant background. Importantly, all volunteers were very supportive of the project, and none raised ethical concerns. After study completion, the volunteers were reinvited to provide feedback on our interpretations of the results.

Reflexivity

Reflexivity is important to highlight how subjectivity may have impacted on research findings [40]. The research team embraces a PR approach in BD. GJ, FL, and SHJ previously developed the POETIC framework for PR in BD. They anticipated that it would capture many aspects shared on the web, but data analysis would reveal new aspects and deeper insights into the experience of PR in everyday life.

Corpus Framework Analysis

Data analysis drew on methods from corpus linguistics [22] and qualitative framework analysis [41], which we call corpus framework analysis. Quantitative corpus-linguistics methods derive frequency lists of the words in the corpus; identify keywords that occur statistically significantly more frequently in the corpus compared to other language samples; and find collocations, that is, words a target word co-occurs with more frequently than by chance. The main qualitative method is to analyze the context of specific words or phrases in so called concordances. Key lemmas in the PR-BD corpus were identified by comparing it to a reference corpus of posts with low similarity to the PR terms list via #LancsBox (version 6.0; Lancaster University) [42]. A lemma is the dictionary form of a word; for example, "recovering" and "recovered" are word forms of the lemma "recover." To focus on the most salient topics of the PR-BD corpus, key lemmas overused at least twice at a significance level of $P<.0001$ [43] and used by at least 5% of users were analyzed. See Document S6 in Multimedia Appendix 1 for methodological details.

The key lemmas were coded into the POETIC framework via concordance analysis. First, overall impressions of all concordances were noted after sorting them according to the lemma, left, and right context (20 words each) in #LancsBox. Subsequently, 30 randomly sampled concordances for each key lemma were coded into the second-level POETIC categories (see the codebook in Appendix B of Jagfeld et al [8]). The coders read the full post if the 40 words did not provide enough context and noted impressions for every key lemma again. Finally, concordances that did not fit into an existing POETIC category were coded inductively. GJ coded all key lemmas, and SHJ and FL audited 6 key lemmas each.

Finally, new key lemmas that were not in the PR terms list and absent PR terms [44] were analyzed. Absence was defined as 0 frequency in the PR-BD corpus or a lower relative frequency than in the reference corpus. Additionally, collocations were analyzed via the #LancsBox *GraphColl* tool for some key lemmas. To do so, content words (noun, verb, adjective, and adverb) within a context of 5 words left and right of the target term and a minimum collocation frequency of 5 were ranked according to cubed mutual information [24].

Results

The S-BiDD data set [45] contains 21,407,595 posts by 19,685 users (available for noncommercial research after signing a data usage agreement). The programming code is publicly available [46].

Coding the PR Relevance of Posts and Constructing the PR-BD Corpus

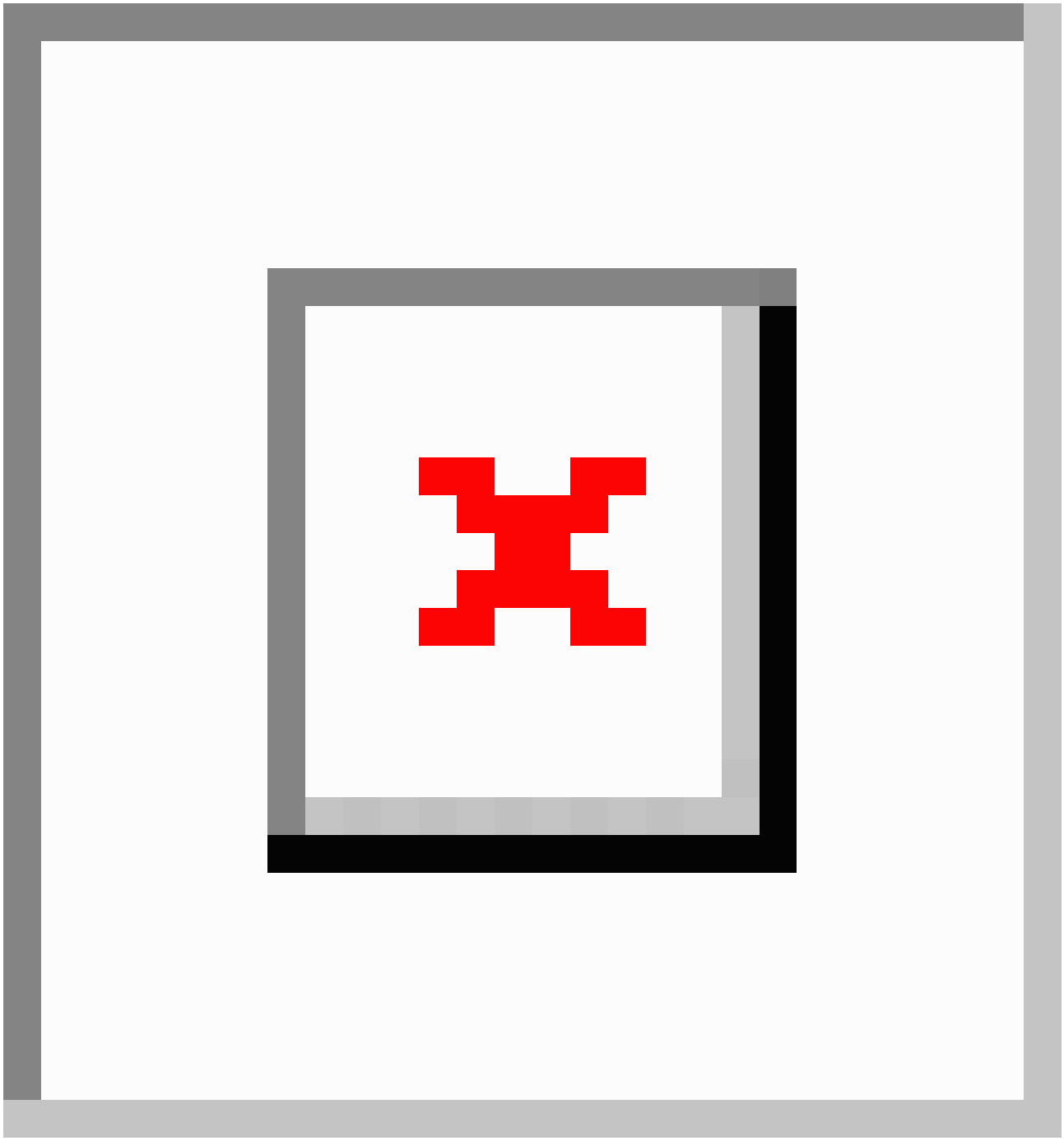
Following a blind trial of coding the PR relevance of 20 posts and a subsequent discussion, GJ and CH achieved moderate agreement (Cohen κ=0.51; 77/100, 77% observed agreement) in coding the remaining 100 posts (see Table S20 in Multimedia Appendix 1). The team resolved all disagreements. In total, 66 (31%) of the 210 posts were coded as PR relevant (see Table S21 in Multimedia Appendix 1). Based on this, the PR-BD corpus comprises posts with a PR score above 0.025 to balance precision in selecting PR-relevant posts and corpus size (see Table S22 in Multimedia Appendix 1). The PR-BD corpus has 4462 posts with 1,337,080 words by 1982 users. The reference corpus of posts with a PR score below 0.013 (see Table S23 in Multimedia Appendix 1) comprises 25,197 posts with 4,700,834 words by 6075 users.

Concordance Analysis With the POETIC Framework

In all, 130 lemmas met the prespecified keyness criteria. Figure 2 shows the domain and category frequencies, extrapolated from the 30 concordance lines coded for each key lemma and color coded according to Tol's [47] light scheme for color-blindness accessibility. Table S28 in Multimedia Appendix 1 lists the key lemmas coded into each category, and Table S29 in Multimedia Appendix 1 lists the categories that each key lemma was coded into. Overall, the POETIC framework captured the experiences in the PR-BD corpus very well: there was evidence for all categories. Only 16% (9303/59,199) of key lemma instances fell into the new "Not POETIC" domain rather than the existing framework. The text below briefly reviews each domain with

key lemmas in italics, highlighting the differences between the original framework and the web-based data. Tables S30 and S31 in Multimedia Appendix 1 provide illustrative quotes for all categories.

Figure 2. Frequency of POETIC domains and categories and new categories. MH: mental health; POETIC: Purpose and Meaning, Optimism and Hope, Empowerment, Tensions, Identity, Connectedness.



Purpose and Meaning

"Purpose and meaning" was the most frequent domain and contained the most salient topic differences between the POETIC review and web-based data. Although participants in the POETIC review mainly discussed the meaningful life and social role of being a *parent*, web-based discussions focused on reproductive decisions. Participants discussed perceived *risks* they might be *responsible* for, for example, the *possibility* for their *child* to *develop* MH issues that affected their *decision* or *choice* to *bear* and *raise kids*. In the "Work" category, extended to include formal education, many discussions focused on *struggles* around *studying* and *graduating college*. No participant in the POETIC review reported financial or housing issues, whereas several web-based users complained about a low *quality* of *life* due to *money* problems, causing homelessness or inability to afford treatment. Spirituality was discussed more frequently and richer than in the POETIC review. Users often wondered whether to regard their experiences as truly spiritual or rather as (hypo)manic symptoms.

Optimism and Hope

Reddit users differed in their "Belief" in the possibility of *recovery*. The mainstream opinion on Reddit was that BD is a "chronic condition that cannot be *cured*, only *managed*." Users questioned whether feeling fully *recovered* was not just a temporary experience caused by (hypo)mania. In the "*Positive* thinking and valuing *success*" category, many users were *grateful* for aspects of their BD experiences; for example, *challenges* provide *opportunities* for *growth* and demonstrating *strength*.

Empowerment

As in the POETIC review, "Self-management and personal *responsibility*" was the most frequent and richest category. Forum users generally considered (taking *steps* towards) *maintaining* a *healthy lifestyle* (including *routines* or *schedules*, *diet*, *exercise*, and *coping skills*) as an individual's *responsibility* to reach *recovery*. In contrast, experiencing MH symptoms or feeling stuck in their recovery was regarded outside of someone's responsibility if they followed professional or mainstream forum advice. The "Controversial role of medication" category included concerns about drug effects on the *baby* during pregnancy or nursing and alternative non–evidence-based treatments such as the keto *diet* or cannabis, which were not present in the POETIC review.

Tensions

Experiences coded in the "Tensions" domain were similar to the POETIC review. Several participants shared feeling more comfortable to discuss "Ambivalence" around (hypo)mania on the web. Some asked if there was a *possibility* to *enjoy* increased *motivation* and *confidence* to make *progress* in their goals without the hypomania getting out of control.

Identity

Some participants shared rich *success* stories in the "Rebuilding positive sense of self" category, in which they moved away from *shame* and *guilt* by *forgiving* themselves for past behaviors and toward *accepting* themselves, whereas others were struggling with this process. *Shame* associated with *stigma* in the *society* was another focus of discussions, and some participants shared creative ways for overcoming stigma.

Connectedness

Regarding "Connectedness," users mainly discussed relationships and *support* from others. Although there were positive accounts, participants often discussed *struggles* with romantic *relationships* or *marriage* and *friendships* and complained about issues with *professional* and *family support*, similar to the POETIC review. However, the web-based accounts, particularly those of relationship and family problems, appeared more candid, for example, with them also discussing *sexual* issues, *trauma*, and *shame*.

Not POETIC

Inductive coding of the 645 concordance lines that did not fit into the POETIC framework revealed that they were unrelated to individuals' PR or lived experience. Most quotes discussed other MH issues without PR relevance (symptoms, genetics and heredity, treatment, diagnosis, societal issues, and scientific research), followed by storytelling of their own or others' situation without PR relevance; direct interactions between forum users, for example, giving advice or congratulating; and discussions of non-MH issues.

New PR Terms

Although 99 (76%) of 130 key lemmas were PR terms, 31 (24%) key lemmas were new. Of these 31 lemmas, 15 (48%) conveyed similar meanings to PR terms; for example, *brother* likened other family members in the PR terms list such as son or nephew. Another 7 (23%) new key lemmas introduced aspects not covered by PR terms. For example, *baby*, *raise*, and *bear* were related to reproductive decision-making; *childhood* was related to making sense of MH issues via early traumatic experiences; and *environment* was related to a focus on structural or societal circumstances rather than the individual (see Table S32 in Multimedia Appendix 1).

Absent PR Terms

Only 13% (n=54) of the 416 unique PR terms (after removing spelling and phraseological variants) were absent: 46 (11%) did not appear in the PR-BD corpus and 8 (2%) were underused compared to the reference corpus. The underused PR terms referred to symptoms (*high mood*, *mania*, *manic*, and *sleep*) or medical MH professionals (*doctor*, *pdoc* [psychiatrist], and *psychiatrist*; see Table S33 in Multimedia Appendix 1). These terms were relevant for some PR domains but also were strongly associated with clinical recovery. All PR terms missing in the PR-BD corpus were also missing in the reference corpus. They were mostly complex phrases, for example, *brush yourself off*, and none indicated aspects that were not covered by other key lemmas (see Table S34 in Multimedia Appendix 1).

Feedback From People With Lived Experience

Two volunteers who had commented on the first exploratory study provided feedback on the main study results. Overall, they valued the results and agreed with our findings but indicated limitations of the data, as reflected in the *Discussion* section. One volunteer argued that categorizations of experiences can be problematic for masking individual differences. Conversely, the other volunteer had found it particularly helpful to align some of her behaviors with CHIME categories because this gave her a sense of being on the right track.

Discussion

This study analyzed Reddit posts of people with a BD diagnosis via corpus framework analysis to learn about the lived experience of PR in BD and validate the POETIC framework.

Key Findings in Relationship to Previous Work

The primary study aim was to provide new insights on PR in BD. Indeed, the web-based data contained candid, in-the-moment experiences that traditional qualitative data collection is unlikely to retrieve. For example, 1 user posted about their experiences in a current manic episode on 2 subsequent days: "Yesterday I posted here about the realization that I've entered a manic episode." Other users shared things on the web that they had not shared elsewhere: 'Talking about

this part of my inner world to a psychiatrist would require a lot of trust for me." The users had different interpretations of elated mood as signs of recovery, spiritual experiences, helpful motivational boosts, or dangerous MH symptoms to avoid. Quantitative [48,49] and qualitative [50,51] evidence shows that web-based anonymity affords personal self-disclosures and discussions of sensitive and stigmatized issues.

The results show that 3 POETIC domains were featured the most in Reddit discussions: "Purpose and meaning" (particularly reproductive decision making, work, and formal education), "Connectedness" (romantic relationships and social support), and "Empowerment" (self-management and personal responsibility). In line with a recent quantitative review [52], the concerns raised on Reddit pointed to a wide range of social and occupational functioning among people with a BD diagnosis: some were not working or leaving their house and therefore sought support on the web, whereas others asked for specific advice to further improve their already functional lifestyle. The popularity of the "Self-management and personal responsibility" category agrees with recent quantitative findings. A review by Mezes et al [53] found positive associations between PR and psychological characteristics focusing on control and personal agency, and a longitudinal study identified positive impacts of adaptive coping and balanced risk-taking on PR [54].

Importantly, the analysis highlighted PR issues that exclusively or more frequently came up on the web. This might be due to differences in sample demographics and data collection methods between this study and those included in the POETIC review. Users in the S-BiDD data set were younger than those in the studies included in the POETIC review: the S-BiDD data set users had a mean age of 32 years versus 45 years in the POETIC review, 30% (5866/19,685) versus 17% (18/163) of participants were aged between 18-29 years, and 7% (1299/19,685) versus 34% (36/163) were aged between 50-64 years [8,30]. This might explain why perspectives on transitioning into adulthood with BD, challenges of college education, and reproductive decision-making exclusively surfaced in the web-based data. Sahota and Sankar [20] summarized their qualitative analysis of discussions of genetic risk and reproductive decision-making in 2 BD subreddits as centering around the manageability of parenting a child for people with a BD diagnosis, which aligns well with the experiences found in this study.

Moreover, users in the S-BiDD data set were overwhelmingly from the United States [30], whereas all POETIC review studies stemmed from countries that provide at least a basic level of free public MH care and social security (the United Kingdom, Norway, Australia, Canada, China, Spain, and Turkey). This may explain why existential financial issues such as (threat of) homelessness and the inability to afford treatment surfaced only in the web-based data. Since health insurance in the United States (except for Medicare for those aged 65+ y) is either employer provided or privately paid, individuals who cannot work due to their MH issues lose their insurance and in turn access to professional support, often causing MH issues to exacerbate, for example, by abruptly stopping medication. One Reddit user described this as a "vicious cycle." It also appears plausible that Reddit users stem from a different socioeconomic group than the participants recruited into the POETIC review studies.

The secondary aim of this study was to validate the POETIC framework. Results confirmed that the framework usefully captured PR experiences shared on the web. Web-based users discussed all second-level POETIC categories, and only 645 of the 3900 analyzed concordance lines could not be accommodated in the framework, demonstrating its comprehensiveness.

Strengths and Limitations

Three aspects of this study constitute both strengths and limitations. First, using online forums as a data source provided rich, candid, and in-the moment experiences. However, there is limited background and demographic information on the online forum users (but see Jagfeld et al [30] for an analysis of these properties in the users in the S-BiDD data set), and they are not representative of the general population with a BD diagnosis. One user in the PR-BD corpus posted "My hunch is that r/bipolarreddit overrepresents those who are struggling, who, understandably, may be more pessimistic about everything." One volunteer shared his experience that discussions on Reddit MH forums mainly followed a mainstream opinion and that deviant opinions were ignored or suppressed. McDonald and Woodward-Kron [27] support this with corpus-linguistics evidence that BD forum users over time shifted from advice seeking to giving and used more medicalized language. Similarly, Vayreda and Antaki [18] showed that established BD forum users urged new members to seek a formal diagnosis and reinforced a biomedical view of BD. Our Reddit study provides one lens on the lived experience of some people that can complement studies of other MH forums and other sources, such as one-on-one interviews.

Second, the list of PR terms facilitated focusing on the concept of interest among the wealth of data, yet it arguably biased the data selection. Nevertheless, 52% (16/31) of the key lemmas that were not PR terms contributed new PR aspects. Moreover, explicitly stating our expectations of PR aspects via the terms list enabled us to identify absent aspects in the data.

Third, corpus-linguistics methods, particularly the coding of key lemmas, allowed the analysis of more data than traditional qualitative methods. However, single words probably more readily capture topic-like (eg, "Relationships") rather than theme-like (eg, "Balancing acceptance with ambitions") categories. Therefore, the relative category frequencies should be interpreted with some caution.

Research Implications

This study has at least 4 research implications. First, it demonstrates the usefulness of analyzing online forum posts to tap into authentic and candid accounts of lived experience of MH issues. Second, this study serves as the first validation of the POETIC framework. Ideally, this encourages other researchers to apply it in their research. Third, the combination of corpus linguistics and qualitative framework analysis allowed the analysis of large amounts of data. Hence, corpus framework analysis may also be useful for future studies of text data, such as therapy transcripts (eg, [55]). Lastly, the S-BiDD data set

and derived corpora are available for future research, for example, on other aspects of the lived experience of BD.

Clinical Implications

This study identifies the key issues relevant to PR in BD shared by people with lived experience on the web and extends previous knowledge from interviews and focus groups. These findings, including the quotes in Tables S30 and S31 in Multimedia Appendix 1, are a rich resource for understanding more about the experience of PR in BD for individuals living with BD, their loved ones and informal carers, and MH professionals. This is also relevant for recent initiatives to educate MH professionals on the lived experience of severe MH issues, such as the current "Understanding psychosis and BD" training for the UK National Health Service [56]. Subsequently, issues identified in this study may provide helpful starting points for therapists to collaboratively consider them with their clients, for example, in recovery-focused therapy [57,58].

Individuals discussed issues on the web that they considered contentious and personal and were not comfortable sharing offline, such as sexuality, spirituality, and (hypo)mania. Recovery-focused therapies that are free to work with whatever model the clients bring for their BD experiences [58] may be particularly suitable to create a therapeutic environment where clients feel comfortable to discuss such sensitive issues. Moreover, Jones et al [59] showed that recovery-focused therapy reduces the positive self-appraisal of hypomanic experiences.

Reproductive decision-making surfaced as a major issue for young adults living with BD, and dedicated counseling on this topic may be advisable. Although understanding genetic vulnerability and risk data in MH is challenging, there is evidence that genetic counseling can offer effective support [60].

Conclusions

This study analyzed 4462 Reddit posts by 1982 people with an S-BiDD within the POETIC framework [8] for PR in BD. It is the first to analyze online forum data on PR. This study confirmed the validity of the POETIC framework to also capture PR experiences shared on the web and highlighted new aspects in PR that did not come up in previous studies using interviews and focus groups. It also demonstrated the utility of integrating corpus linguistics and qualitative framework analysis to identify key themes within large text data sets. By indicating the key areas that people focus on when posting freely, this study provides rich insights into the lived experience of PR in BD for formal and informal carers of people with a BD diagnosis.

Acknowledgments

The authors wish to thank the volunteers with lived experience who provided helpful insights for conducting and writing up this research. GJ is grateful to Enrica Troiano for insightful discussions on designing and evaluating the coding for personal recovery relevance. She would also like to thank Gavin Brookes and the participants of the 6th Corpora & Discourse International Conference for helpful comments on the corpus construction and analysis. The authors also thank Matthew Coole for testing the code release. This study was completed as part of a PhD studentship for GJ, which was funded by the Faculty of Health and Medicine at Lancaster University, United Kingdom.

Authors' Contributions

GJ led on the study design, supported by SHJ, FL, and PR. GJ collected the data, conducted the analyses, and drafted the manuscript. SHJ and FL audited the analyses (personal recovery relevance of posts and framework analysis of concordance lines). CH double coded the personal recovery relevance of posts and reviewed the paraphrasing of selected quotes. All coauthors discussed and agreed on the results, commented on the draft manuscript, and approved the final version. SHJ, FL, and PR obtained the funding for this study.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary reports and documents.
[DOCX File, 513 KB - medinform_v11i1e46544_app1.docx]

References

1. Jones SH, Lobban F, Cook A, Division of Clinical Psychology. Understanding Bipolar Disorder: Why Some People Experience Extreme Mood States and What Can Help: The British Psychological Society; 2010. [doi: 10.53841/bpsrep.2010.rep151]
2. Rowland TA, Marwaha S. Epidemiology and risk factors for bipolar disorder. Ther Adv Psychopharmacol 2018 Sep;8(9):251-269. [doi: 10.1177/2045125318769235] [Medline: 30181867]
3. Michalak EE, Yatham LN, Lam RW. Quality of life in bipolar disorder: a review of the literature. Health Qual Life Outcomes 2005 Nov 15;3:72. [doi: 10.1186/1477-7525-3-72] [Medline: 16288650]

4. Novick DM, Swartz HA, Frank E. Suicide attempts in bipolar I and bipolar II disorder: a review and meta-analysis of the evidence. Bipolar Disord 2010 Feb;12(1):1-9. [doi: 10.1111/j.1399-5618.2009.00786.x] [Medline: 20148862]
5. Anthony WA. Recovery from mental illness: the guiding vision of the mental health service system in the 1990s. Psychosoc Rehabil J 1993;16(4):11-23. [doi: 10.1037/h0095655]
6. Murray G, Leitan ND, Thomas N, et al. Towards recovery-oriented Psychosocial interventions for bipolar disorder: quality of life outcomes, stage-sensitive treatments, and Mindfulness mechanisms. Clin Psychol Rev 2017 Mar;52:148-163 S0272-7358(16)30220-3. [doi: 10.1016/j.cpr.2017.01.002] [Medline: 28129636]
7. van Weeghel J, van Zelst C, Boertien D, Hasson-Ohayon I. Conceptualizations, assessments, and implications of personal recovery in mental illness: a scoping review of systematic reviews and meta-analyses. Psychiatr Rehabil J 2019 Jun;42(2):169-181. [doi: 10.1037/prj0000356] [Medline: 30843721]
8. Jagfeld G, Lobban F, Marshall P, Jones SH. Personal recovery in bipolar disorder: systematic review and "best fit" framework synthesis of qualitative evidence - a POETIC adaptation of CHIME. J Affect Disord 2021 Sep 1;292:375-385. [doi: 10.1016/j.jad.2021.05.051] [Medline: 34139411]
9. Leamy M, Bird V, Le Boutillier C, Williams J, Slade M. Conceptual framework for personal recovery in mental health: systematic review and narrative synthesis. Br J Psychiatry 2011 Dec;199(6):445-452. [doi: 10.1192/bjp.bp.110.083733] [Medline: 22130746]
10. Stuart SR, Tansey L, Quayle E. What we talk about when we talk about recovery: a systematic review and best-fit framework synthesis of qualitative literature. J Ment Health 2017 Jun;26(3):291-304. [doi: 10.1080/09638237.2016.1222056] [Medline: 27649767]
11. Slade M, Leamy M, Bacon F, et al. International differences in understanding recovery: systematic review. Epidemiol Psychiatr Sci 2012 Dec;21(4):353-364. [doi: 10.1017/S2045796012000133] [Medline: 22794507]
12. Briggs CL. Learning How to Ask: A Sociolinguistic Appraisal of the Role of the Interview in Social Science Research: Cambridge University Press; 1986. [doi: 10.1017/CBO9781139165990]
13. Janetzko D. Nonreactive data collection online. In: Fielding NG, Lee RM, Blank G, editors. The SAGE Handbook of Online Research Methods: SAGE Publications Ltd; 2016:76-91. [doi: 10.4135/9781473957992]
14. Seale C, Charteris-Black J, MacFarlane A, McPherson A. Interviews and Internet forums: a comparison of two sources of qualitative data. Qual Health Res 2010 May;20(5):595-606. [doi: 10.1177/1049732309354094] [Medline: 20008955]
15. Kramer ADI, Fussell SR, Setlock LD. Text analysis as a tool for analyzing conversation in online support groups. Presented at: CHI EA '04: CHI '04 Extended Abstracts on Human Factors in Computing Systems; Apr 24-29, 2004; Vienna Austria p. 1485-1488. [doi: 10.1145/985921.986096]
16. Gkotsis G, Oellrich A, Hubbard T, et al. The language of mental health problems in social media. Presented at: Proceedings of the Third Workshop on Computational Lingusitics and Clinical Psychology; Jun 16, 2016; San Diego, CA, USA p. 63-73. [doi: 10.18653/v1/W16-0307]
17. Coppersmith G, Dredze M, Harman C, Hollingshead K. From ADHD to SAD: analyzing the language of mental health on Twitter through self-reported diagnoses. Presented at: Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology; Jun 5, 2015; Denver, Colorado p. 1-10. [doi: 10.3115/v1/W15-1201]
18. Vayreda A, Antaki C. Social support and unsolicited advice in a bipolar disorder online forum. Qual Health Res 2009 Jul;19(7):931-942. [doi: 10.1177/1049732309338952] [Medline: 19556400]
19. Mandla A, Billings J, Moncrieff J. "Being bipolar": a qualitative analysis of the experience of bipolar disorder as described in internet blogs. Issues Ment Health Nurs 2017 Oct;38(10):858-864. [doi: 10.1080/01612840.2017.1355947] [Medline: 28872998]
20. Sahota PKC, Sankar PL. Bipolar disorder, genetic risk, and reproductive decision-making: a qualitative study of social media discussion boards. Qual Health Res 2020 Jan;30(2):293-302. [doi: 10.1177/1049732319867670] [Medline: 31409193]
21. Bauer R, Bauer M, Spiessl H, Kagerbauer T. Cyber-support: an analysis of online self-help forums (online self-help forums in bipolar disorder). Nord J Psychiatry 2013 Jun;67(3):185-190. [doi: 10.3109/08039488.2012.700734] [Medline: 22817138]
22. McEnery T, Hardie A. Corpus Linguistics: Method, Theory and Practice: Cambridge University Press; 2011. [doi: 10.1017/CBO9780511981395]
23. Semino E, Demjén Z, Hardie A, Payne S, Rayson P. Cancer and the End of Life: A Corpus-Based Study: Routledge, Taylor & Francis; 2017. [doi: 10.4324/9781315629834]
24. Hunt D, Brookes G. Corpus, Discourse and Mental Health: Bloomsbury Academic; 2020. [doi: 10.5040/9781350059207]
25. Baker P. Using Corpora in Discourse Analysis: Continuum; 2006. [doi: 10.5040/9781350933996]
26. Abdo M, Ali A, Sarhan N. Analyzing judgment in bipolar depression patients' narratives using syntactic patterns: a corpus-based study. Egyptian Journal of Language Engineering 2019 Apr 1;6(1):1-11. [doi: 10.21608/ejle.2019.59103]
27. McDonald D, Woodward-Kron R. Member roles and identities in online support groups: perspectives from corpus and systemic functional linguistics. Discourse & Communication 2016 Apr;10(2):157-175. [doi: 10.1177/1750481315615985]
28. Reddit. URL: www.reddit.com/ [accessed 2023-10-02]
29. Reddit. Alexa. 2022. URL: www.alexa.com/siteinfo/reddit.com [accessed 2020-02-27]

30. Jagfeld G, Lobban F, Rayson P, Jones SH. Understanding who uses Reddit: profiling individuals with a self-reported bipolar disorder diagnosis. Presented at: Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology; Jun 11, 2021; Online p. 1-14. [doi: 10.18653/v1/2021.clpsych-1.1]
31. McEnery T, Xiao R, Tono Y. Corpus-Based Language Studies: An Advanced Resource Book: Routledge, Taylor & Francis; 2006.
32. Jagfeld G. Subreddit topics. GitHub. 2022. URL: github.com/glorisonne/reddit_bd_mood_posting_mh/blob/main/data/subreddit_topics.csv [accessed 2023-02-23]
33. Jagfeld G. List of bipolar disorder terms. GitHub. 2021. URL: github.com/glorisonne/reddit_bd_user_characteristics/blob/master/disclosure-patterns/condition-terms/bipolar-filter-terms.txt [accessed 2023-02-23]
34. Jagfeld G, Lobban F, Humphreys C, Rayson P, Jones SH. List of personal recovery terms. GitHub. 2022. URL: github.com/glorisonne/reddit_bd_recovery/blob/main/data/PR_terms.csv [accessed 2023-02-23]
35. Spärck Jones K. A statistical interpretation of term specificity and its application in retrieval. J Doc 2004 Oct 1;60(5):493-502. [doi: 10.1108/00220410410560573]
36. Robertson S. Understanding inverse document frequency: on theoretical arguments for IDF. J Doc 2004 Oct;60(5):503-520. [doi: 10.1108/00220410410560582]
37. Research Board. Ethics guidelines for internet-mediated research. The British Psychological Society. 2021 Jun 7. URL: www.bps.org.uk/guideline/ethics-guidelines-internet-mediated-research [accessed 2023-10-02]
38. Morant N, Chilman N, Lloyd-Evans B, Wackett J, Johnson S. Acceptability of using social media content in mental health research: a reflection. comment on “Twitter users” views on mental health crisis resolution team care compared with stakeholder interviews and focus groups: qualitative analysis. JMIR Ment Health 2021 Aug 17;8(8):e32475. [doi: 10.2196/32475] [Medline: 34402799]
39. People in Research. URL: www.peopleinresearch.org/ [accessed 2023-10-02]
40. Finlay L, Gough G, editors. Reflexivity: A Practical Guide for Researchers in Health and Social Sciences: John Wiley & Sons; 2008.
41. Pope C, Ziebland S, Mays N. Qualitative research in health care. analysing qualitative data. BMJ 2000 Jan 8;320(7227):114-116. [doi: 10.1136/bmj.320.7227.114] [Medline: 10625273]
42. Brezina V, Platt W, McEnery T. #Lancsbox v. 6.0. Lancaster University. 2021. URL: corpora.lancs.ac.uk/lancsbox [accessed 2023-10-02]
43. Rayson P, Berridge D, Francis B. Extending the Cochran rule for the comparison of word frequencies between Corpora. Presented at: 7th International Conference on Statistical Analysis of Textual Data (JADT 2004); Mar 10-12, 2004; Louvain-la-Neuve, Belgium p. 926-936.
44. Duguid A, Partington A. Absence: you don’t know what you’re missing. Or do you? In: Taylor C, Marchi A, editors. Corpus Approaches to Discourse: A Critical Review: Routledge, Taylor & Francis; 2018:38-59. [doi: 10.4324/9781315179346]
45. Jagfeld G, Jones SH, Rayson P, Lobban F. Self-reported BD diagnosis (S-BiDD) data set. Lancaster University. 2022. URL: 10.17635/lancaster/researchdata/589 [accessed 2023-10-02]
46. Jagfeld G. Code accompanying the paper “how people with a bipolar disorder diagnosis talk about personal recovery in peer online support forums: corpus framework analysis using POETIC.”. GitHub. 2022. URL: github.com/glorisonne/reddit_bd_recovery/ [accessed 2023-02-23]
47. Tol P. Colour schemes. Netherlands Institute for Space Research. 2021 Aug 18. URL: personal.sron.nl/~pault/data/colourschemes.pdf [accessed 2023-10-02]
48. De Choudhury M, De S. Mental health discourse on Reddit: self-disclosure, social support, and anonymity. Presented at: Eighth International AAAI Conference on Weblogs and Social Media; Jun 1-4, 2014; Ann Arbor, MI p. 71-80. [doi: 10.1609/icwsm.v8i1.14526]
49. Pavalanathan U, de Choudhury M. Identity management and mental health discourse in social media. Proc Int World Wide Web Conf 2015 May;2015(Companion):315-321. [doi: 10.1145/2740908.2743049] [Medline: 27376158]
50. Wright K. Communication in health-related online social support groups/communities: a review of research on predictors of participation, applications of social support theory, and health outcomes. Review of Communication Research 2016;4:65-87. [doi: 10.12840/issn.2255-4165.2016.04.01.010]
51. Smith-Merry J, Goggin G, Campbell A, McKenzie K, Ridout B, Baylosis C. Social connection and online engagement: insights from interviews with users of a mental health online forum. JMIR Ment Health 2019 Mar 26;6(3):e11084. [doi: 10.2196/11084] [Medline: 30912760]
52. Akers N, Lobban F, Hilton C, Panagaki K, Jones SH. Measuring social and occupational functioning of people with bipolar disorder: a systematic review. Clin Psychol Rev 2019 Dec;74. [doi: 10.1016/j.cpr.2019.101782] [Medline: 31751878]
53. Mezes B, Lobban F, Costain D, et al. Recovery beyond clinical improvement - recovery outcomes measured for people with bipolar disorder between 1980 and 2020. J Affect Disord 2022 Jul 15;309:375-392 S0165-0327(22)00412-8. [doi: 10.1016/j.jad.2022.04.075] [Medline: 35469910]
54. Mezes B, Lobban F, Costain D, Longson D, Jones SH. Psychological factors in personal and clinical recovery in bipolar disorder. J Affect Disord 2021 Feb 1;280(Pt A):326-337. [doi: 10.1016/j.jad.2020.11.044] [Medline: 33221719]

55. Tay D, Qiu H. Modeling linguistic (a)synchrony: a case study of therapist-client interaction. Front Psychol 2022 May 23;13:903227. [doi: 10.3389/fpsyg.2022.903227] [Medline: 35677134]
56. Understanding psychosis and bipolar disorder. Health Education England. 2020 Mar. URL: www.hee.nhs.uk/sites/default/files/documents/Understanding psychosis and bipolar disorder.pdf [accessed 2023-10-02]
57. Tyler E, Lobban F, Sutton C, et al. A pilot randomised controlled trial to assess the feasibility and acceptability of recovery-focused therapy for older adults with bipolar disorder. BJPsych Open 2022 Oct 24;8(6):e191. [doi: 10.1192/bjo.2022.582] [Medline: 36278451]
58. Jones SH, Smith G, Mulligan LD, et al. Recovery-focused cognitive-behavioural therapy for recent-onset bipolar disorder: randomised controlled pilot trial. Br J Psychiatry 2015 Jan;206(1):58-66. [doi: 10.1192/bjp.bp.113.141259] [Medline: 25213157]
59. Jones SH, Knowles D, Howarth E, Lobban F, Emsley R. Mediation analysis of recovery-focused therapy for recent-onset bipolar disorder. J Affect Disord Reports 2021 Jul;5:100175. [doi: 10.1016/j.jadr.2021.100175]
60. Hippman C, Ringrose A, Inglis A, et al. A pilot randomized clinical trial evaluating the impact of genetic counseling for serious mental illnesses. J Clin Psychiatry 2016 Feb;77(2):e190-e198. [doi: 10.4088/JCP.14m09710] [Medline: 26930535]

Abbreviations

BD: bipolar disorder
CHIME: Connectedness, Hope and Optimism, Identity, Meaning and Purpose, Empowerment
MH: mental health
NLP: natural language processing
POETIC: Purpose and Meaning, Optimism and Hope, Empowerment, Tensions, Identity, Connectedness
PR: personal recovery
S-BiDD: self-reported bipolar disorder diagnosis

Edited by C Lovis; submitted 23.02.23; peer-reviewed by A Molinard-Chenu, E Jaafar; revised version received 12.09.23; accepted 13.09.23; published 08.11.23.

Please cite as:
Jagfeld G, Lobban F, Humphreys C, Rayson P, Jones SH
How People With a Bipolar Disorder Diagnosis Talk About Personal Recovery in Peer Online Support Forums: Corpus Framework Analysis Using the POETIC Framework
JMIR Med Inform 2023;11:e46544
URL: https://medinform.jmir.org/2023/1/e46544
doi:10.2196/46544

Detection of Suicidal Ideation in Clinical Interviews for Depression Using Natural Language Processing and Machine Learning: Cross-Sectional Study

Tim M H Li[1], PhD; Jie Chen[1], MD, PhD; Framenia O C Law[1], BA; Chun-Tung Li[1], PhD; Ngan Yin Chan[1], PhD; Joey W Y Chan[1], MBChB; Steven W H Chau[1], MBBS, MSc; Yaping Liu[1], MD, PhD; Shirley Xin Li[2,3], PhD, DClinPsy; Jihui Zhang[1,4], MD, PhD; Kwong-Sak Leung[5,6], PhD; Yun-Kwok Wing[1], MBChB

1

2

3

4

5

6

Corresponding Author:
Tim M H Li, PhD

Abstract

Background: Assessing patients' suicide risk is challenging, especially among those who deny suicidal ideation. Primary care providers have poor agreement in screening suicide risk. Patients' speech may provide more objective, language-based clues about their underlying suicidal ideation. Text analysis to detect suicide risk in depression is lacking in the literature.

Objective: This study aimed to determine whether suicidal ideation can be detected via language features in clinical interviews for depression using natural language processing (NLP) and machine learning (ML).

Methods: This cross-sectional study recruited 305 participants between October 2020 and May 2022 (mean age 53.0, SD 11.77 years; female: n=176, 57%), of which 197 had lifetime depression and 108 were healthy. This study was part of ongoing research on characterizing depression with a case-control design. In this study, 236 participants were nonsuicidal, while 56 and 13 had low and high suicide risks, respectively. The structured interview guide for the Hamilton Depression Rating Scale (HAMD) was adopted to assess suicide risk and depression severity. Suicide risk was clinician rated based on a suicide-related question (H11). The interviews were transcribed and the words in participants' verbal responses were translated into psychologically meaningful categories using Linguistic Inquiry and Word Count (LIWC).

Results: Ordinal logistic regression revealed significant suicide-related language features in participants' responses to the HAMD questions. Increased use of anger words when talking about work and activities posed the highest suicide risk (odds ratio [OR] 2.91, 95% CI 1.22-8.55; *P*=.02). Random forest models demonstrated that text analysis of the direct responses to H11 was effective in identifying individuals with high suicide risk (AUC 0.76-0.89; *P*<.001) and detecting suicide risk in general, including both low and high suicide risk (AUC 0.83-0.92; *P*<.001). More importantly, suicide risk can be detected with satisfactory performance even without patients' disclosure of suicidal ideation. Based on the response to the question on hypochondriasis, ML models were trained to identify individuals with high suicide risk (AUC 0.76; *P*<.001).

Conclusions: This study examined the perspective of using NLP and ML to analyze the texts from clinical interviews for suicidality detection, which has the potential to provide more accurate and specific markers for suicidal ideation detection. The findings may pave the way for developing high-performance assessment of suicide risk for automated detection, including online chatbot-based interviews for universal screening.

(JMIR Med Inform 2023;11:e50221) doi:10.2196/50221

KEYWORDS

depression; suicidal ideation; clinical interview; machine learning; natural language processing; automated detection

Introduction

Up to 77% of individuals who died by suicide had contact with their primary care provider within 12 months prior to their death [1]. Suicidal ideation is a significant risk factor for suicidal death and is an important clinical concern [2-4]. Screening for suicidal ideation is a standard practice in health care settings. However, a past study reported poor agreement in rating suicide

risk among primary care providers [5]. Assessing patients' suicide risk is challenging, especially among those who deny suicidal ideation and perceive suicide as taboo to talk about [6-8]. In recent years, researchers have attempted to screen suicide risk unobtrusively based on implicit language-based clues in patients' speech [9,10]. A study applied a naive Bayes classifier to patients' verbal responses in clinical interviews to detect suicide risk with an area under the curve (AUC) of 0.63 [9]. This line of research is supported by recent behavioral and neuroimaging studies that demonstrated a close relationship between language use and social-emotional processing [11].

A systematic review has suggested that first-person singular pronouns and negative emotion words are language features of people who are suicidal [12]. However, first-person singular pronouns and negative emotion words are also found among those who are prone to depression [13,14], and first-person singular pronoun use has been proposed as a specific language marker of depression in a meta-analysis [15]. This result is concordant with the idea of depressive self-focus and rumination on past events, particularly on negative memories in depression. Given the overlapping language features among patients with depression and those with suicide risk, there is a need to look for language features that are specific in predicting suicide among patients with depression. Previous studies reported other linguistic features (eg, prepositions and verbs) for predicting suicide risk [12,16]. Another study differentiating suicide notes from depression notes and neutral blog posts found adverbs, cognitive processing words, and death words as the most significant language features of suicidality [17].

Natural language processing (NLP) and machine learning (ML) have been used to detect suicide risk [18,19]. Previous studies have incorporated NLP and ML into suicide detection for universal screening on social media [20-22]. Researchers have identified explicit suicide expressions (eg, suicide notes) [23] and texts written by suicidal individuals (eg, songs, poems, diaries) [24], but these have mostly been searched for in social media posts and other personal documents [25]. Few studies have evaluated the techniques in clinical settings [18,19,26]. This study used a text analysis approach using NLP and ML techniques for suicidal ideation detection from the words spoken by participants in clinical interviews for depression. It is hypothesized that (1) language features extracted from responses to interview questions are associated with suicidal ideation; (2) there is a difference in language features between suicidal ideation and depression; and (3) the model can contribute to more accurate suicidal ideation detection, especially among patients with depression.

Methods

Recruitment

This cross-sectional study was part of an ongoing digital phenotyping research project in characterizing depression with a case-control design [27]. Patients with lifetime major depressive disorder (MDD) were recruited from outpatient clinics in a local university-affiliated hospital. The diagnosis of any psychiatric disorder was made by the attending psychiatrist. Controls were recruited from the community. The Mini-International Neuropsychiatric Interview, version 5.0, was used to check if they had any *Diagnostic and Statistical Manual of Mental Disorders*, fourth edition, diagnosis [28]. We found that 8% (9/117) of the community sample had lifetime MDD, and these individuals were considered as cases. The diagnostic criteria for MDD included having (1) 5 or more depressive symptoms for ≥2 weeks, (2) either depressed mood or loss of interest and pleasure, (3) symptoms causing significant distress or impairment, and (4) no manic or hypomanic behavior. Participants who, at some point in their lives, had ever received a diagnosis were classified as cases of lifetime MDD. Inclusion criteria for participation were (1) being a native Cantonese speaker and (2) being a Chinese adult aged 18 to 65 years. Participants (1) with any voice, speech, and language problems; (2) with any history of psychiatric disorder other than MDD; and (3) who were incompetent to give written informed consent were excluded. The data collection was conducted between October 2020 and May 2022. This study included 197 cases and 108 controls (n=305).

Ethical Considerations

Ethical approval was obtained from the Joint Chinese University of Hong Kong–New Territories East Cluster clinical research ethics committee (2020.492).

Clinical Measurements

The structured interview guide for the Hamilton Depression Rating Scale (HAMD) was adopted [29]. The HAMD-17 shows high interrater reliability [30]. The interviews were conducted by the second author (JC), who holds an MD and a PhD. Each interview took 15 to 30 minutes. We sequentially probed 14 questions on the HAMD-17 (H1 to H14) to assess depression symptoms (we rated questions H15 to H17 based on observation during the interview). Depression severity was classified as no depression (score 0-7), mild depression (8-16), and moderate to severe depression (≥17) [29].

H11, which was used to assess suicide risk, asks "Since last week, have you had any thoughts that life is not worth living?" Suicide risk was rated in five progressive levels: (1) having no suicidal thoughts; (2) feeling life is not worth living; (3) having wishes to be dead, or any thoughts of possible death of self; (4) having suicidal ideation or gestures; and (5) having attempts at suicide. JC and TMHL rated H11 with high interrater reliability (κ=0.92). Discrepancies in ratings were discussed until a consensus was reached. A total of 236 of 305 participants (77%) who did not report any suicidal thoughts were classified as the nonsuicidal group. The severity of suicidal ideation ranged from passive ideation (ie, having wishes to be dead) to active ideation (ie, having suicidal thoughts and behaviors); persons with active ideation are more prone to suicide attempts and deaths than those with passive ideation [3]. Thus, 33 and 23 participants who felt life was not worth living and had wished to be dead, respectively, were classified as the low-suicide-risk group (18%); 12 participants and 1 participant who had suicidal thoughts and acting out behavior, respectively, were classified as the high-suicide-risk group (4%). Multimedia Appendix 1 includes verbatim quotations of the verbal responses to H11.

Feature Extraction

The interviews were recorded and transcribed by a research assistant with a psychology background. The transcripts were checked by the first author (TMHL). As interword spacing is absent in Chinese texts (eg, "I want to kill myself" in Chinese would become "Iwanttokillmyself"), Chinese word segmentation was needed to separate words. For Chinese word segmentation, the study used a deep learning–based Chinese word segmentation engine, fastHan, which included local text samples for training and testing its segmentation model, achieving over 90% agreement with human segmentation [31].

Words were translated into psychologically meaningful categories using Linguistic Inquiry and Word Count (LIWC) [32]. There are 71 categories in the Chinese version of LIWC [33]. To investigate if a language feature f existed within a verbal response, we calculated the proportion of words r_i in the response that matched with any of the words c_j listed in an LIWC category using the following formula:

where $R=\{r_1, r_2, \ldots\}$ and $C=\{c_1, c_2, \ldots\}$ denotes the collection of words in the response and the LIWC category, respectively, while $m(r, c)$ represents checking for an exact match between r and c (which returns 1=matched or 0=not matched), and $|R|$ denotes the number of words in the response. All the features were calculated by the proportion of words of each category relative to text length (as a percentage). Using the relative frequency minimized the confounding factor of text length in interview responses.

Statistical Analysis

All analyses were conducted using R (version 4.2.0; R Foundation for Statistical Computing). A P value <.05 was considered statistically significant. Descriptive statistics for continuous variables are shown as means and SDs, while categorical variables are presented as numbers and percentages. Age and the number of words in interview responses were compared among the nonsuicidal, low-suicide-risk, and high-suicide-risk groups using a 1-way ANOVA, while gender and depression severity were compared with the chi-square test among the groups. The Cramér V was used to measure the associations of clinician ratings of the HAMD questions with suicide risk.

Ordinal logistic regression was performed to model the relationship between language features (predictors) and suicide risk (the outcome)—ordered in three progressive levels from (1) nonsuicidal, (2) low suicide risk, to (3) high suicide risk [34]. The analysis was conducted for the responses to H11 (a suicide-related question) and the responses to other questions (content without suicide disclosure). The regression models were adjusted for age, gender, and depression severity. Adjusted odds ratios (ORs) with 95% CIs were calculated as a measure of the strength of association. The ORs were standardized across the HAMD questions and visualized using a heatmap. Euclidean distance was used as the similarity measure for clustering the questions with similar language features associated with suicide risk. Bonferroni correction for multiple testing was applied.

Random forest, an ML classification technique [35], was used to detect participants with suicide risk, including (1) high suicide risk and (2) any suicide risk (both low and high risk) among (1) all participants, (2) participants with lifetime MDD, (3) participants with lifetime MDD and unremitted depression (HAMD-17 score ≥8), (4) participants with lifetime MDD and remitted depression, and (5) control participants, based on language features extracted from their interview responses. All classification results were evaluated by leave-one-out cross-validation. Receiver operating characteristic curve analysis was used for analyzing the accuracy of classification results. Statistics included the AUC and 95% CI, sensitivity, and specificity of the ML classifiers. For each classifier, sensitivity and specificity at the optimal cutoff were computed.

Results

Characteristics of the Participants

Table 1 shows the characteristics of the 3 groups, namely the nonsuicidal, low-suicide-risk, and high-suicide-risk groups. No significant age or gender differences were found among the 3 groups. Suicide risk was correlated with depression severity (V=0.45, 95% CI 0.36-0.52; P<.001). All participants in the high-suicide-risk group were depressed, of which the majority (8/13, 62%) were moderately to severely depressed. In the low-suicide-risk group, 86% (48/56) had depression. The majority (28/56, 50%) were experiencing mild depression. Only 23% (54/236) of nonsuicidal participants had depression, most of whom (46/236, 20%) had mild depression. Participants, on average, generated 387.93 (SD 349.19) words in response to all 14 interview questions. There were significant differences among the 3 groups in terms of the number of words in their interview responses. The low-suicide-risk group generated longer responses than the nonsuicidal group (P<.001) and high-suicide-risk group (P=.02). The high-suicide-risk group also uttered marginally more words compared to the nonsuicidal group (P=.05).

Table . Characteristics of the participants (N=305).

Characteristics		Nonsuicidal (n=236)	Low suicide risk (n=56)	High suicide risk (n=13)	*P* value
Age (years), mean (SD)		52.63 (11.57)	54.46 (11.88)	51.69 (15.42)	.27
Gender, n (%)					.37
	Female	132 (56)	37 (66)	7 (54)	
	Male	104 (44)	19 (34)	6 (46)	
Lifetime major depressive disorder, n (%)					<.001
	Yes	133 (56)	51 (91)	13 (100)	
	No	103 (44)	5 (8)	0 (0)	
Depression severity[a], n (%)					<.001
	None	182 (77)	8 (14)	0 (0)	
	Mild	46 (20)	28 (50)	5 (38)	
	Moderate or severe	8 (3)	20 (36)	8 (62)	
Number of words in responses, mean (SD)					
	To all questions except H11[b]	313.86 (280.60)	604.57 (434.45)	420.08 (187.05)	<.001
	To H11 only	7.91 (16.78)	48.77 (58.96)	25.62 (18.89)	<.001

[a]Severity ranges for the Hamilton Depression Rating Scale–17: no depression (0-7); mild depression (8-16); moderate to severe depression (≥17).

[b]H11: Hamilton Depression Rating Scale question 11.

Table 2 shows the associations of clinician ratings of the HAMD questions with suicide risk. All the ratings were associated with suicide risk (*V*=0.14-0.40). The clinician rating of H12 (on anxiety psychic) had the strongest association with suicide risk, whereas the rating of H8 (on insomnia late) had the weakest association with suicide risk.

Table . Associations of clinician ratings of the Hamilton Depression Rating Scale (HAMD) questions with suicide risk (based on H11).

HAMD item	Question type	Cramér *V* (95% CI)	*P* value
H1	Depressed mood	0.36 (0.26-0.42)	<.001
H2	Work and activities	0.36 (0.27-0.43)	<.001
H3	Genital symptoms	0.21 (0.12-0.28)	<.001
H4	Somatic symptoms gastrointestinal	0.30 (0.21-0.37)	<.001
H5	Loss of weight	0.18 (0.08-0.25)	.005
H6	Insomnia early	0.23 (0.13-0.30)	<.001
H7	Insomnia middle	0.17 (0.07-0.24)	.003
H8	Insomnia late	0.14 (0.01-0.20)	.03
H9	Somatic symptoms general	0.30 (0.21-0.37)	<.001
H10	Feelings of guilt	0.36 (0.26-0.43)	<.001
H12	Anxiety psychic	0.40 (0.30-0.46)	<.001
H13	Anxiety somatic	0.31 (0.22-0.38)	<.001
H14	Hypochondriasis	0.28 (0.18-0.34)	<.001

Associations of Suicide Risk With Language Features in H11

Suicide risk was assessed by clinicians based on H11 (a suicide-related question). Table 3 shows the significant associations of suicide risk with language features in verbal responses to H11. Suicide risk was positively associated with linguistic (ie, function words, verbs, adverbs, prepositions, and numbers), psychological (ie, social, cognitive, and biological process words, relativity words, death words, and fillers), and Chinese-specific categories (ie, postpositions, quantity units, multifunction words, and tense markers). After adjusting the analyses for depression severity, 11 language features remained significant. Among the 11 features, increased use of past tense markers posed the highest suicide risk. For every 1% increase in past tense markers, the odds of being more likely to have

higher suicide risk (low or high suicide risk vs nonsuicidal) were multiplied 1.24 times (OR 1.24, 95% CI 1.09-1.43; *P*=.002).

Table . Significant associations of suicide risk with language features in verbal responses to H11 (Hamilton Depression Rating Scale question 11) using ordinal logistic regression.

LIWC[a] category	Nonsuicidal (n=236), mean (SD)	Low suicide risk (n=56), mean (SD)	High suicide risk (n=13), mean (SD)	Odds ratio (95% CI)[b]	*P* value	Odds ratio (95% CI)[c]	*P* value
Function words	19.48 (22.32)	42.48 (15.75)	44.25 (17.64)	1.05 (1.03-1.07)	<.001	1.04 (1.02-1.06)	<.001
Verbs	3.30 (10.62)	15.08 (10.11)	21.90 (8.03)	1.13 (1.09-1.16)	<.001	1.08 (1.05-1.11)	<.001
Auxiliary verbs	0.39 (1.68)	4.14 (7.46)	5.27 (4.12)	1.28 (1.18-1.40)	<.001	1.17 (1.08-1.28)	<.001
Adverbs	7.10 (14.73)	11.59 (9.06)	11.06 (8.14)	1.02 (1.00-1.04)	.03	1.02 (1.00-1.05)	.07
Prepositions	0.92 (2.85)	7.07 (7.52)	12.22 (10.04)	1.24 (1.18-1.31)	<.001	1.19 (1.13-1.26)	<.001
Numbers	0.32 (2.48)	1.70 (4.72)	0.48 (1.37)	1.09 (1.01-1.18)	.02	1.03 (0.94-1.11)	.45
Postpositions	0.86 (3.64)	3.15 (5.31)	2.22 (3.58)	1.10 (1.04-1.16)	.002	1.06 (1.00-1.13)	.06
Quantity units	1.06 (4.04)	3.05 (3.56)	1.53 (2.54)	1.09 (1.02-1.16)	.009	1.04 (0.96-1.12)	.32
Multifunction words	1.98 (9.79)	8.75 (9.70)	11.11 (8.70)	1.09 (1.05-1.13)	<.001	1.05 (1.03-1.09)	<.001
Tense markers	0.30 (1.87)	1.52 (2.88)	3.09 (4.27)	1.30 (1.16-1.46)	<.001	1.15 (1.03-1.29)	.01
Past tense markers	0.15 (1.25)	1.09 (2.45)	2.43 (3.99)	1.40 (1.22-1.63)	<.001	1.24 (1.09-1.43)	.002
Present tense markers	0.00 (0.00)	0.23 (0.94)	0.00 (0.00)	1.71 (1.08-2.90)	.02	1.07 (0.65-1.78)	.80
Social processes	0.60 (2.35)	1.74 (3.12)	0.39 (0.95)	1.10 (1.00-1.20)	.04	0.93 (0.82-1.04)	.22
Family	0.02 (0.26)	0.37 (1.10)	0.00 (0.00)	1.66 (1.15-2.41)	.006	1.04 (0.70-1.54)	.86
Cognitive processes	13.89 (19.00)	20.12 (14.96)	18.05 (12.42)	1.02 (1.00-1.03)	.02	1.01 (0.99-1.03)	.16
Discrepancy	0.94 (4.62)	4.33 (7.42)	5.74 (4.38)	1.12 (1.06-1.19)	<.001	1.08 (1.03-1.14)	<.001
Tentative	1.16 (3.96)	5.41 (7.76)	2.96 (2.82)	1.10 (1.06-1.16)	<.001	1.05 (1.00-1.11)	.04
Biological processes	0.14 (0.92)	1.01 (2.36)	1.04 (3.10)	1.34 (1.16-1.58)	<.001	1.04 (0.87-1.23)	.69
Body	0.09 (0.64)	0.45 (1.15)	0.83 (2.35)	1.55 (1.21-2.02)	<.001	1.04 (0.77-1.41)	.79
Health	0.05 (0.37)	0.56 (2.03)	0.21 (0.77)	1.32 (1.08-1.74)	.01	1.03 (0.83-1.32)	.76
Relativity	2.23 (6.13)	9.21 (9.77)	7.23 (7.26)	1.09 (1.06-1.13)	<.001	1.05 (1.02-1.09)	.004
Motion	0.73 (3.47)	2.14 (3.18)	1.09 (1.79)	1.08 (1.01-1.16)	.03	1.03 (0.93-1.11)	.55
Space	0.88 (3.64)	2.38 (3.41)	3.85 (4.41)	1.12 (1.04-1.21)	.002	1.07 (0.99-1.15)	.07
Time	1.12 (3.87)	6.22 (9.16)	4.03 (5.72)	1.10 (1.06-1.15)	<.001	1.06 (1.01-1.11)	.01
Death	0.11 (1.32)	0.68 (1.89)	0.83 (2.35)	1.27 (1.06-1.55)	.02	1.12 (0.93-1.30)	.18
Filler	0.71 (2.64)	1.57 (2.80)	1.57 (3.31)	1.10 (1.01-1.20)	.03	1.01 (0.90-1.12)	.88

[a]LIWC: Linguistic Inquiry and Word Count.

[b]Adjusted for age and gender.

[c]Adjusted for age, gender, and depression severity.

Associations of Suicide Risk With Language Features Across Other HAMD Questions

Suicide-related language features varied across other HAMD questions (see Multimedia Appendix 2). Among the language features, increased use of anger words in H2 posed the highest suicide risk. For every 1% increase in anger words, the odds of being more likely to have suicide risk were multiplied 2.91 times (OR 2.91, 95% CI 1.22-8.55; *P*=.02). Some language features, on the other hand, were negatively associated with suicide risk. For example, increased use of feeling words (OR 0.36, 95% CI 0.12-0.82; *P*=.04), future tense markers (OR 0.50, 95% CI 0.23-0.84; *P*=.03), and anxiety words (OR 0.57, 95% CI 0.29-0.88; *P*=.04) when responding to H2, H8, and H13, respectively, reduced suicide risk. Overall, six language features—namely function words, auxiliary verbs, prepositions,

multifunction words, cognitive process words, and discrepancy words—were found to be associated with suicide risk in 3 or more questions; these are illustrated in Figure 1.

Figure 1. Significant suicide-related features of language vary across the Hamilton Depression Rating Scale questions H1-14. Six language features (namely function words, auxiliary verbs, prepositions, multifunction words, cognitive process words, and discrepancy words) were found to be associated with suicide risk in 3 or more questions. LIWC: Linguistic Inquiry and Word Count. *P<.05; **P<.01; ***P<.001.

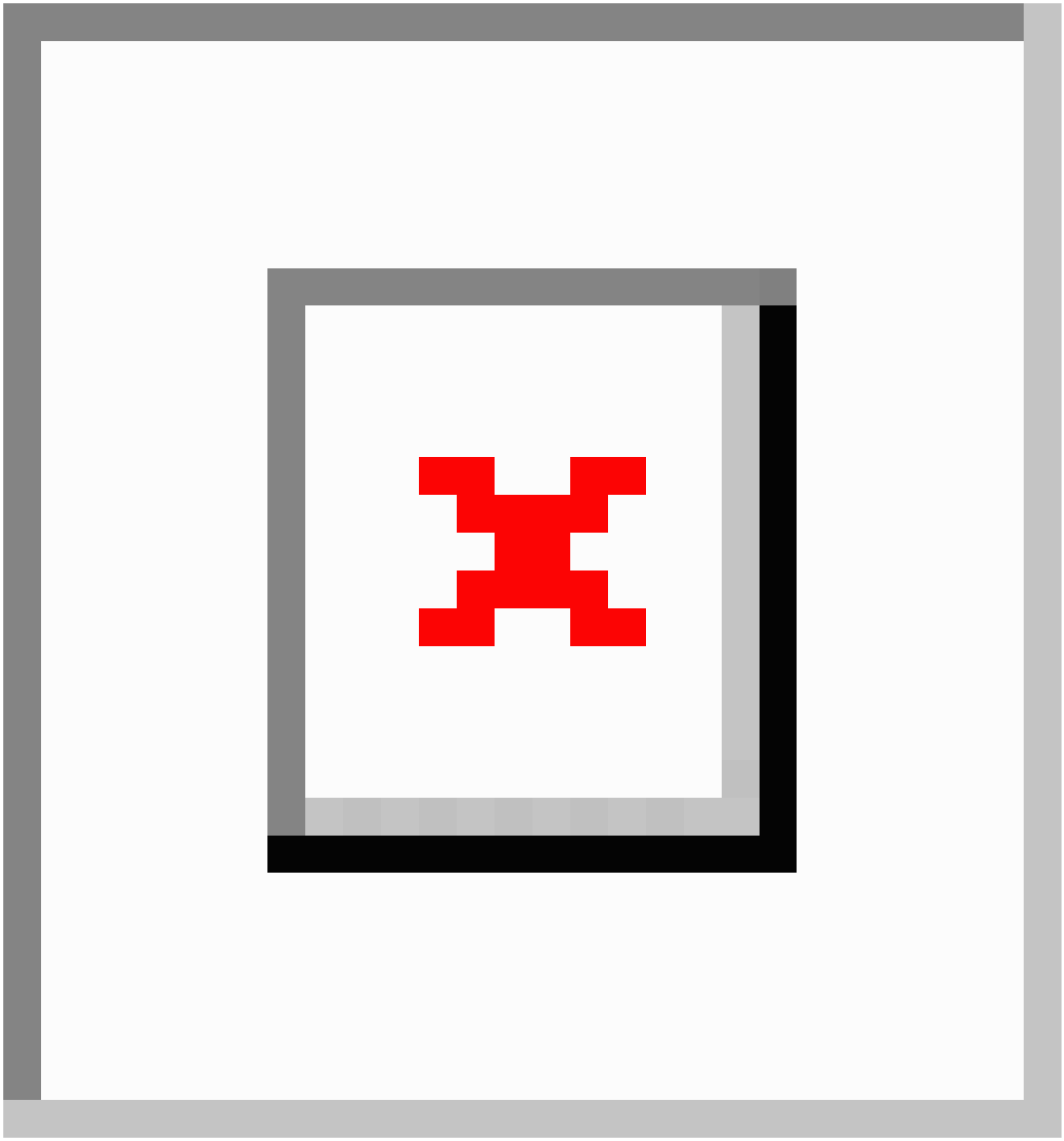
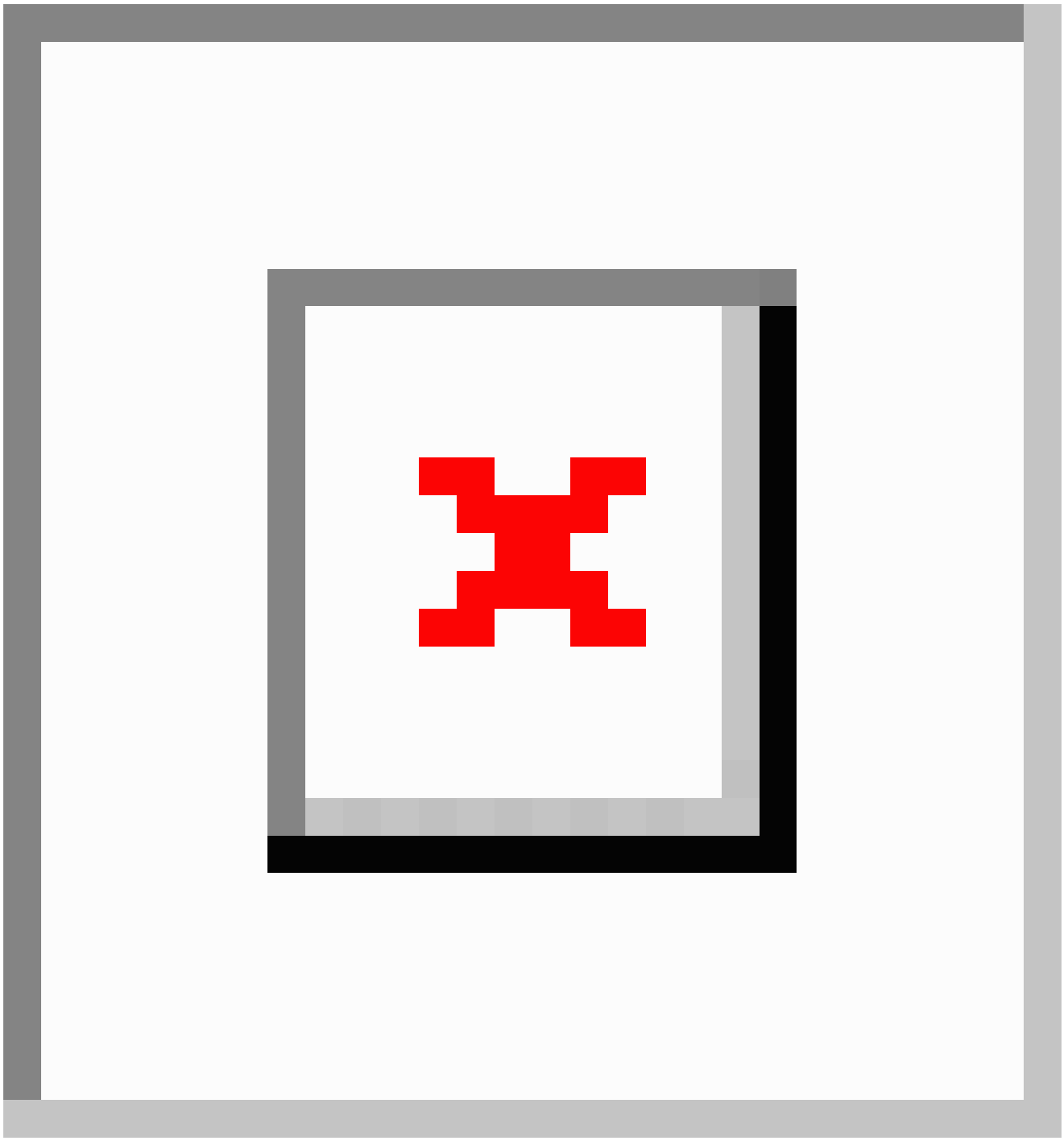


Figure 2 illustrates the hierarchical relationship between the HAMD questions based on the association of language features with suicide risk. Questions with similar suicide-related language features are closer and joined as clusters in the dendrogram. For instance, H10 (asking about feelings of guilt) joined together with H11 (a suicide-related question) as a cluster. While the heights reflect the similarity between the clusters, the largest difference between clusters is between the clusters of H10 and H11 vs the clusters of the other questions. Some possible clusters were observed in the dendrogram, including (1) the suicide-related cluster: H10 and H11; (2) the mood cluster: H2, H8, H1, and H12; (3) the appetite cluster: H5, H4, and H13; (4) the somatic symptom cluster: H7, H9, and H14; and (5) the sexual activity cluster: H3 and H6.

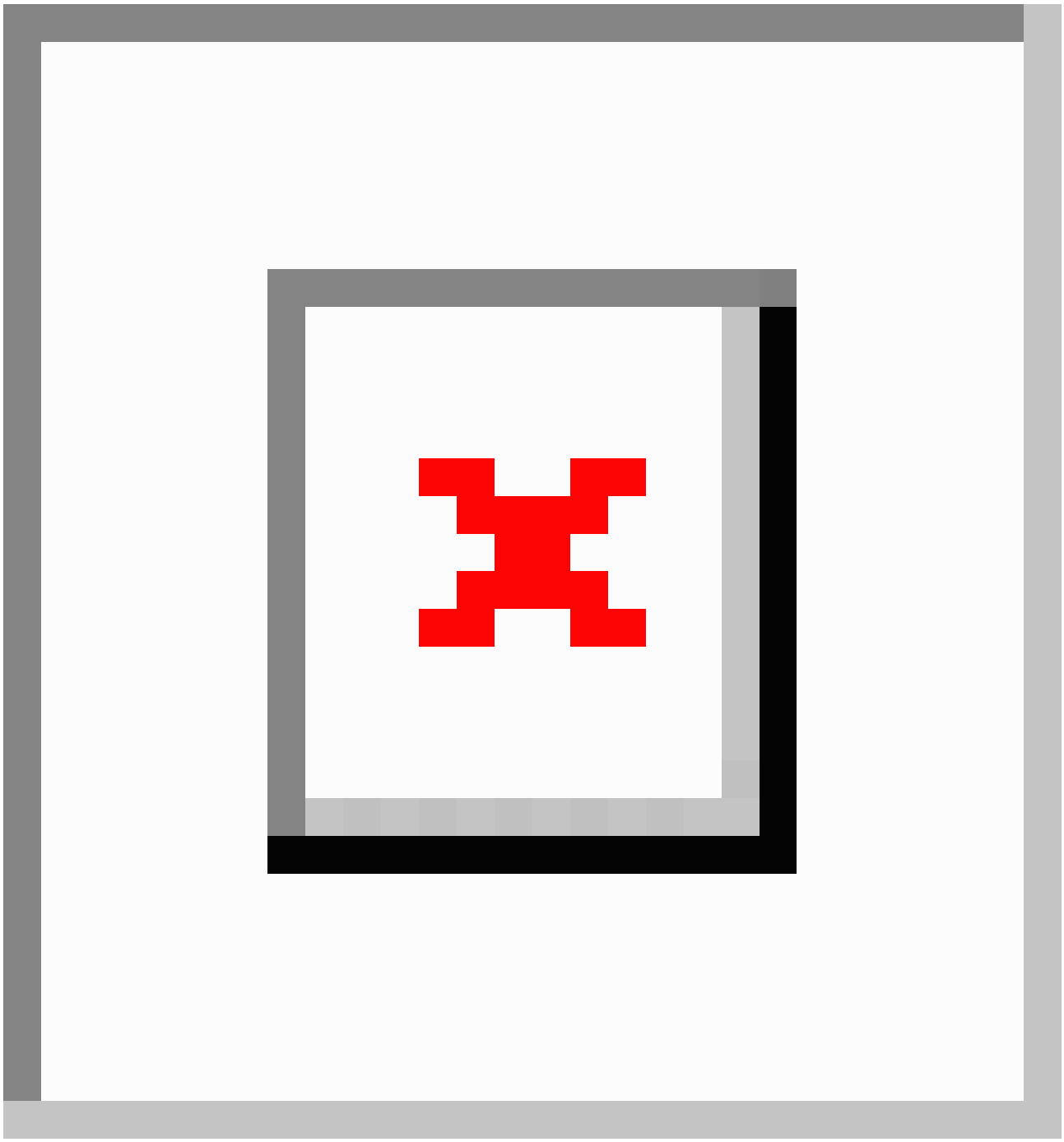
Figure 2. Heatmap of associations between language features and suicide risk across the Hamilton Depression Rating Scale (HAMD) questions (H1-14, where H11 asks about suicidal ideation) in all participants (n=305). The heatmap shows the standardized odds ratios (darker colors represent greater associations with suicide risk) obtained from ordinal logistic regression models adjusted for age, gender, and depression severity. Euclidean distance was used as the similarity measure for clustering the HAMD questions with similar language features associated with suicide risk shown in the dendrogram. LIWC: Linguistic Inquiry and Word Count.



The Performance of Suicide Risk Detection

Figure 3 depicts the performance for detecting suicide risk based on verbal responses to the HAMD questions in the ML analysis. Based on the responses to H11 (the outcome measure), ML models were trained to identify individuals with high risk in (Figure 3A) all participants, (Figure 3B) those with lifetime MDD, and (Figure 3C) those with lifetime MDD and unremitted depression (AUC 0.76-0.89; *P*<.001; sensitivity=0.69-0.85; specificity=0.73-0.84; Multimedia Appendix 3). H11 could also be used to detect suicide risk in general (including both high and low suicide risk) in (Figure 3D) all participants, (Figure 3E) those with lifetime MDD, and (Figure 3F) those with lifetime MDD and unremitted depression (AUC 0.83-0.92; *P*<.001; sensitivity=0.77-0.84; specificity=0.78-0.84). None of the control participants (without lifetime MDD) had high suicide risk, while 3 healthy controls reported low suicide risk. Using only H11, suicide risk could be detected among control participants (AUC 0.70; *P*=.04; sensitivity=0.80; specificity=0.71).

Figure 3. The performance, in terms of area under the curve (AUC), for detecting (A-C) active and (D-F) active and passive suicidal ideation based on participants' verbal responses to the Hamilton Depression Rating Scale (HAMD) questions H1 to H14 (where H11 asks about suicidal ideation) using random forest with leave-one-out cross-validation. Participants with active ideation (n=13) were detected among (A) all participants (n=305), (B) those with lifetime major depressive disorder (MDD; n=197), and (C) those with lifetime MDD and HAMD-17 score ≥8 (n=109). Participants with active or passive ideation were detected (n=69, 64, and 60, respectively) among (D) all participants, (E) those with lifetime MDD, and (F) those with lifetime MDD and HAMD-17 score ≥8. The error bars represent the 95% CI for AUC. *P<.05; **P<.01; ***P<.001.



Based on the responses to H14 (on hypochondriasis), ML models were trained to identify individuals with high suicide risk in (Figure 3A) all participants (AUC 0.76; P<.001; sensitivity=0.69; specificity=0.70), (Figure 3B) those with lifetime MDD (AUC 0.76; P<.001; sensitivity=0.77; specificity=0.77), and (Figure 3C) those with lifetime MDD and unremitted depression (AUC 0.75; P=.003; sensitivity=0.69; specificity=0.69). H4, H6, H8, and H10 could also be used to detect high suicide risk. Except for H7 and H8, other questions could be used to detect suicide risk in general (including both high and low suicide risk) in (Figure 3D) all participants. Based on the responses to H1, H4, H5, H9, H10, and H12, ML models identified individuals with suicide risk in (Figure 3E) those with lifetime MDD.

Discussion

Overview

This study aimed to provide a novel perspective on suicidal ideation detection by analyzing texts in clinician-administered,

structured interviews with NLP and ML. LIWC extracts human-understandable language features without knowing the interview content and can preserve patient privacy in clinical research. The study complements previous research with social media data (over 80% of the related studies were conducted using social media texts [19]) by (1) incorporating clinician-rated measures as the outcome, (2) investigating texts generated for specific purposes (ie, responding to the HAMD questions), and (3) exploring the added value of text analysis to clinical practice. The findings show that significant language features extracted from verbal responses to interview questions are associated with clinician-rated suicide risk. There is a difference in language features between suicidal ideation and depression. The ML models demonstrate that using direct responses to H11 is effective in identifying participants with suicide risk. More importantly, suicide risk can also be detected with satisfactory performance even without patients' disclosure of suicidal ideation.

Principal Results

The study distinguished suicidal ideation from its major confounding effect, depression, with an aim to investigate solely suicide-related language features. The use of past tense markers, verbs, and prepositions in response to a suicide-related question posed a higher risk than other language features. This finding on past tense markers is coherent with previous studies in which rumination occurred in suicidal individuals or past suicide attempts were described [12,15]. An increase in verbs is also found in previous research, which implies an aggravated suicide risk when suicidal intentions and thoughts become actions [14,16]. Prepositions are a relatively new language feature finding. This is consistent with a systematic review that analyzed 75 studies, only one of which reported increased use of prepositions of suicidal thoughts and behaviors [12]. In the modern Chinese language, neuroscience evidence suggests that prepositions are probably not a separate word class from verbs (the action words in a sentence) [36], which are also more frequently used by suicidal people. Another explanation for prepositions as connectors of words is that they demand and convey information about location, time, or direction [37]. An increase in prepositions use by suicidal people when responding to a suicide-related question may highlight the existence of concrete suicide plans.

Many previous depression and suicide detection studies were conducted on social media rather than clinical interviews [18,19]. Social media texts freely generated by users are often nonspecific to suicide or depression [22]. This study analyzed participants' responses to the HAMD questions, specifically focusing on a series of depressive symptomatologies. Results were mixed for suicide-related language features in responses to different questions. Previous studies reported a significant difference in the use of first-person singular pronouns and negative emotion words between suicidal and nonsuicidal groups [12,24]. This study consistently found these patterns posed a higher suicide risk: increased use of anger words on the work and activities question (H2), negative emotion words on the middle insomnia question (H7), and sadness words on the somatic symptoms question (H9). However, the increased use of first-person singular pronouns and anxiety words when responding to H5 (on the loss of weight) and H13 (on anxiety somatic), respectively, were intriguingly found to reduce suicide risk, which seems contradictory to previous studies that found these 2 categories were major language features of suicidal ideation [12]. It is speculated that rather than being generic, suicide-related language features appear to be more topic specific. The use of first-person singular pronouns to describe feelings and daily activities could reflect self-focus, a low level of social integration, and even suicidal rumination [12]. Paying more attention to oneself in the context of body weight management (in H5), on the other hand, may indicate positive self-concept and self-compassion [38]. Using more anxiety words (in H13) to describe the relationship between somatic symptoms and anxiety may signify a good understanding of one's own condition.

ML on the response to H11 (the outcome measure) is, understandably, highly effective in identifying both clinical and control participants with suicide risk, which is consistent with positive findings from previous studies detecting suicide-related social media posts labeled using human annotation [21,23]. With AUCs up to 92%, automatic detection of suicidal ideation in clinical settings, especially among busy and primary care clinics, seems possible. Real-time feedback from the ML models during clinical interviews can potentially facilitate early detection of suicidal ideation [8]. The ML models can be developed with other automated techniques, including chatbot-based interviews for future screening [10]. Furthermore, suicide risk can be detected with satisfactory performance without participants' disclosure of suicidal ideation or attempts (eg, in H10 on feelings of guilt and H14 on hypochondriasis), which provides new evidence on the association of language use with suicidality. In particular, while the clinician rating of H14 is not the strongest predictor for suicide risk, the responses to H14, surprisingly, generate the most powerful language clues to identify high suicide risk in the ML analysis. Verbal responses tend to provide subtle yet more objective language-based information about underlying suicidal ideation, which transcends depressive symptomatologies. The findings suggest that the investigation of suicide-related language features should be an important research topic apart from studying the clinical correlates of suicidality. The evidence strengthens the use of NLP and ML for suicidal ideation detection using implicit language-based clues, especially when health care personnel cannot identify patients' suicide risk upon a direct suicide-related question, or when patients refuse to answer a suicide-related question.

Limitations

First, although we had a relatively large sample size of patients, there were few high-suicide-risk participants. The sample size of at-risk groups will be increased for a more balanced sample. Second, this study focused on a cross-sectional time frame to investigate the discriminating power of language features for suicide detection. A longitudinal study could contribute to a longer time-frame exploration of the temporal association of language use with suicidal thoughts and behaviors. Third, although the HAMD-17 showed high interrater reliability [30] and high intermethod reliability [39], and JC and TMHL rated H11 with high interrater reliability (κ=0.92), the reliability of

the interview ratings of the rest of the HAMD questions in the current study was not assessed. H11 was used to determine whether the participants had suicide risk, while the text analysis was also conducted on the patient's verbal responses in the interview. It is also difficult to avoid errors arising from participants' willful denial of suicidal ideation or participants' belief that talking about suicide is taboo. Future research may need a more comprehensive measure for the outcome of suicidality that is based on more objective benchmarks (eg, actual suicidal behaviors or history). Fourth, the current study did not compare various machine learning algorithms (eg, support vector machines, naive Bayes classifiers, and neural networks), model parameters, and feature sets to achieve the best results. The performance of suicidal ideation detection might be an underestimate. The paper also proposes future directions to use deep learning–based models, such as Bidirectional Encoder Representations from Transformers (BERT), MentalBERT, and other large language models [40] to conduct comparative experiments. Finally, the language and cultural differences between the Chinese and English languages, which could create deviations among the studies, should be investigated. Chinese people seem more conservative in expressing emotions, and topics like suicide may receive media airtime but not be present in their personal lives [6,7]. Only rarely do Chinese or, more broadly, Asian people express personal feelings and suicidal ideation, as they commonly see these as signs of a weak personality (eg, irresponsible, fragile, impulsive, and attention-seeking), which culminates and intertwines with the social stigma of mental disorders in the general population [41].

Conclusions

The study investigated a novel perspective of using NLP and ML to analyze the texts from clinical interviews for suicidal ideation detection, which has the potential to provide more accurate and specific markers for suicide detection. Suicide risk detection is crucial groundwork for precrisis management to respond to safety concerns in automated screening and assessment. Other media such as chatbots and text-based detection have been used in many mental health applications. We hope that the enhanced suicide detection described in this study will augment and strengthen the performance of screening and assessment in the future.

Acknowledgments

This work was supported by the Health and Medical Research Fund (09203066) and a Chinese University of Hong Kong Direct Grant for Research (2021.055 and 2022.073).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Verbatim transcription of the responses to H11 among the 3 groups.
[DOCX File, 39 KB - medinform_v11i1e50221_app1.docx]

Multimedia Appendix 2
Significant associations of suicide risk with language features in verbal responses to the Hamilton Depression Rating Scale (HAMD) questions using ordinal logistic regression.
[DOCX File, 24 KB - medinform_v11i1e50221_app2.docx]

Multimedia Appendix 3
The performance of detecting suicide risk based on verbal responses to the Hamilton Depression Rating Scale (HAMD) questions using random forest with leave-one-out cross-validation.
[DOCX File, 22 KB - medinform_v11i1e50221_app3.docx]

References

1. Mann JJ, Michel CA, Auerbach RP. Improving suicide prevention through evidence-based strategies: a systematic review. Am J Psychiatry 2021 Jul;178(7):611-624. [doi: 10.1176/appi.ajp.2020.20060864] [Medline: 33596680]
2. Suicide worldwide in 2019: global health estimates. World Health Organization. 2021. URL: www.who.int/publications/i/item/9789240026643 [accessed 2023-10-20]
3. Liu RT, Bettis AH, Burke TA. Characterizing the phenomenology of passive suicidal ideation: a systematic review and meta-analysis of its prevalence, psychiatric comorbidity, correlates, and comparisons with active suicidal ideation. Psychol Med 2020 Feb;50(3):367-383. [doi: 10.1017/S003329171900391X] [Medline: 31907085]
4. Hubers AAM, Moaddine S, Peersmann SHM, et al. Suicidal ideation and subsequent completed suicide in both psychiatric and non-psychiatric populations: a meta-analysis. Epidemiol Psychiatr Sci 2018 Apr;27(2):186-198. [doi: 10.1017/S2045796016001049] [Medline: 27989254]

5. Saini P, While D, Chantler K, Windfuhr K, Kapur N. Assessment and management of suicide risk in primary care. Crisis 2014;35(6):415-425. [doi: 10.1027/0227-5910/a000277] [Medline: 25234744]
6. Li A, Huang X, Jiao D, O'Dea B, Zhu T, Christensen H. An analysis of stigma and suicide literacy in responses to suicides broadcast on social media. Asia Pac Psychiatry 2018 Mar;10(1). [doi: 10.1111/appy.12314] [Medline: 29383880]
7. Chen SS, Lam TP, Lam KF, et al. Youths' attitudes toward open discussion of suicide, preferred contexts, and the impact of internet use: an exploratory sequential mixed-methods study in Hong Kong. Int J Soc Psychiatry 2023 May;69(3):575-586. [doi: 10.1177/00207640221123394] [Medline: 36120996]
8. Blades CA, Stritzke WGK, Page AC, Brown JD. The benefits and risks of asking research participants about suicide: a meta-analysis of the impact of exposure to suicide-related content. Clin Psychol Rev 2018 Aug;64:1-12. [doi: 10.1016/j.cpr.2018.07.001] [Medline: 30014862]
9. Shin D, Kim K, Lee SB, et al. Detection of depression and suicide risk based on text from clinical interviews using machine learning: possibility of a new objective diagnostic marker. Front Psychiatry 2022;13:801301. [doi: 10.3389/fpsyt.2022.801301] [Medline: 35686182]
10. Hungerbuehler I, Daley K, Cavanagh K, Garcia Claro H, Kapps M. Chatbot-based assessment of employees' mental health: design process and pilot implementation. JMIR Form Res 2021 Apr 21;5(4):e21678. [doi: 10.2196/21678] [Medline: 33881403]
11. Brooks JA, Shablack H, Gendron M, Satpute AB, Parrish MH, Lindquist KA. The role of language in the experience and perception of emotion: a neuroimaging meta-analysis. Soc Cogn Affect Neurosci 2017 Feb 1;12(2):169-183. [doi: 10.1093/scan/nsw121] [Medline: 27539864]
12. Homan S, Gabi M, Klee N, et al. Linguistic features of suicidal thoughts and behaviors: a systematic review. Clin Psychol Rev 2022 Jul;95:102161. [doi: 10.1016/j.cpr.2022.102161] [Medline: 35636131]
13. Tackman AM, Sbarra DA, Carey AL, et al. Depression, negative emotionality, and self-referential language: a multi-lab, multi-measure, and multi-language-task research synthesis. J Pers Soc Psychol 2019 May;116(5):817-834. [doi: 10.1037/pspp0000187] [Medline: 29504797]
14. Havigerová JM, Haviger J, Kučera D, Hoffmannová P. Text-based detection of the risk of depression. Front Psychol 2019;10:513. [doi: 10.3389/fpsyg.2019.00513] [Medline: 30936845]
15. Edwards T, Holtzman NS. A meta-analysis of correlations between depression and first person singular pronoun use. J Res Pers 2017 Jun;68:63-68. [doi: 10.1016/j.jrp.2017.02.005]
16. Fernández-Cabana M, Jiménez-Féliz J, Alves-Pérez MT, Mateos R, Gómez-Reino Rodríguez I, García-Caballero A. Linguistic analysis of suicide notes in Spain. Eur J Psychiat 2015;29(2):145-155. [doi: 10.4321/S0213-61632015000200006]
17. Schoene AM, Turner AP, De Mel G, Dethlefs N. Hierarchical multiscale recurrent neural networks for detecting suicide notes. IEEE Trans Affect Comput 2021 Feb 5;14(1):153-164. [doi: 10.1109/TAFFC.2021.3057105]
18. Le Glaz A, Haralambous Y, Kim-Dufor DH, et al. Machine learning and natural language processing in mental health: systematic review. J Med Internet Res 2021 May 4;23(5):e15708. [doi: 10.2196/15708] [Medline: 33944788]
19. Zhang T, Schoene AM, Ji S, Ananiadou S. Natural language processing applied to mental illness detection: a narrative review. NPJ Digit Med 2022 Apr 8;5(1):46. [doi: 10.1038/s41746-022-00589-7] [Medline: 35396451]
20. Cheng Q, Li TM, Kwok CL, Zhu T, Yip PS. Assessing suicide risk and emotional distress in Chinese social media: a text mining and machine learning study. J Med Internet Res 2017 Jul 10;19(7):e243. [doi: 10.2196/jmir.7276] [Medline: 28694239]
21. Chau M, Li TMH, Wong PWC, Xu JJ, Yip PSF, Chen H. Finding people with emotional distress in online social media: a design combining machine learning and rule-based classification. MISQ 2020 Jun 1;44(2):933-955. [doi: 10.25300/MISQ/2020/14110]
22. Kelley SW, Mhaonaigh CN, Burke L, Whelan R, Gillan CM. Machine learning of language use on Twitter reveals weak and non-specific predictions. NPJ Digit Med 2022 Mar 25;5(1):35. [doi: 10.1038/s41746-022-00576-y] [Medline: 35338248]
23. Roy A, Nikolitch K, McGinn R, Jinah S, Klement W, Kaminsky ZA. A machine learning approach predicts future risk to suicidal ideation from social media data. NPJ Digit Med 2020;3(1):78. [doi: 10.1038/s41746-020-0287-6] [Medline: 32509975]
24. Li TMH, Chau M, Yip PSF, Wong PWC. Temporal and computerized psycholinguistic analysis of the blog of a Chinese adolescent suicide. Crisis 2014 May 1;35(3):168-175. [doi: 10.1027/0227-5910/a000248]
25. Cheng Q, Lui CSM. Applying text mining methods to suicide research. Suicide Life Threat Behav 2021 Feb;51(1):137-147. [doi: 10.1111/sltb.12680] [Medline: 33624867]
26. Corcoran CM, Cecchi GA. Using language processing and speech analysis for the identification of psychosis and other disorders. Biol Psychiatry Cogn Neurosci Neuroimaging 2020 Aug;5(8):770-779. [doi: 10.1016/j.bpsc.2020.06.004] [Medline: 32771179]
27. Chen J, Li CT, Li TMH, et al. A forgotten sign of depression–the omega sign and its implication. Asian J Psychiatr 2023 Feb;80:103345. [doi: 10.1016/j.ajp.2022.103345] [Medline: 36423435]
28. Sheehan DV, Lecrubier Y, Sheehan KH, et al. The Mini-International Neuropsychiatric Interview (MINI): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. J Clin Psychiatry 1998;59 Suppl 20(20):22-33. [Medline: 9881538]

29. Zimmerman M, Martinez JH, Young D, Chelminski I, Dalrymple K. Severity classification on the Hamilton Depression Rating Scale. J Affect Disord 2013 Sep 5;150(2):384-388. [doi: 10.1016/j.jad.2013.04.028] [Medline: 23759278]
30. Morriss R, Leese M, Chatwin J, Baldwin D, THREAD Study Group. Inter-rater reliability of the Hamilton Depression Rating Scale as a diagnostic and outcome measure of depression in primary care. J Affect Disord 2008 Dec;111(2-3):204-213. [doi: 10.1016/j.jad.2008.02.013] [Medline: 18374987]
31. Geng Z, Yan H, Qiu X, Huang XJ. Fasthan: a BERT-based multi-task Toolkit for Chinese NLP. Presented at: The 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing; Aug 1-6, 2021; Online p. 99-106. [doi: 10.18653/v1/2021.acl-demo.12]
32. Tausczik YR, Pennebaker JW. The psychological meaning of words: LIWC and computerized text analysis methods. J Lang Soc Psychol 2010 Mar;29(1):24-54. [doi: 10.1177/0261927X09351676]
33. Huang CL, Chung CK, Hui N, et al. The development of the Chinese linguistic inquiry and word count dictionary. Chin J Psychol 2012;54(2):185-201.
34. Agresti A. Categorical Data Analysis: John Wiley & Sons; 2002. [doi: 10.1002/0471249688]
35. Liaw A, Wiener M. Classification and regression by randomForest. R News 2002;2(3):18-22 [FREE Full text]
36. Fang HH, Zhang RP, Fang HF, Gao MY, Zheng M, Sun XY. Dependent mechanism of Chinese prepositions processing in the brain: evidence from event-related potentials. Neurosci Bull 2007 Sep;23(5):282-286. [doi: 10.1007/s12264-007-0042-x] [Medline: 17952137]
37. Pennebaker JW, Mehl MR, Niederhoffer KG. Psychological aspects of natural language use: our words, our selves. Annu Rev Psychol 2003;54(1):547-577. [doi: 10.1146/annurev.psych.54.101601.145041] [Medline: 12185209]
38. Ziemer KS, Lamphere BR, Raque-Bogdan TL, Schmidt CK. A randomized controlled study of writing interventions on college women's positive body image. Mindfulness 2019 Jan;10(1):66-77. [doi: 10.1007/s12671-018-0947-7]
39. Yung HY, Yeung WT, Law CW. The reliability of symptom assessment by telepsychiatry compared with face to face psychiatric interviews. Psychiatry Res 2022 Oct;316:114728. [doi: 10.1016/j.psychres.2022.114728] [Medline: 35908348]
40. Greco CM, Simeri A, Tagarelli A, Zumpano E. Transformer-based language models for mental health issues: a survey. Pattern Recognit Lett 2023 Mar;167:204-211. [doi: 10.1016/j.patrec.2023.02.016]
41. Zou W, Tang L, Bie B. The stigmatization of suicide: a study of stories told by college students in China. Death Stud 2022;46(9):2035-2045. [doi: 10.1080/07481187.2021.1958396] [Medline: 34323165]

Abbreviations

AUC: area under the curve
BERT: Bidirectional Encoder Representations from Transformers
HAMD: Hamilton Depression Rating Scale
LIWC: Linguistic Inquiry and Word Count
MDD: major depressive disorder
ML: machine learning
NLP: natural language processing
OR: odds ratio

Edited by C Lovis; submitted 23.06.23; peer-reviewed by T Zhang, Y Bai; revised version received 31.07.23; accepted 23.08.23; published 01.12.23.

Please cite as:
Li TMH, Chen J, Law FOC, Li CT, Chan NY, Chan JWY, Chau SWH, Liu Y, Li SX, Zhang J, Leung KS, Wing YK
Detection of Suicidal Ideation in Clinical Interviews for Depression Using Natural Language Processing and Machine Learning: Cross-Sectional Study
JMIR Med Inform 2023;11:e50221
URL: https://medinform.jmir.org/2023/1/e50221
doi:10.2196/50221

Identifying Risk Factors Associated With Lower Back Pain in Electronic Medical Record Free Text: Deep Learning Approach Using Clinical Note Annotations

Aman Jaiswal[1], BTech; Alan Katz[2], MBChB, MSc; Marcello Nesca[2], BCom, BA, MSc; Evangelos Milios[1], EE, SM, PhD

1

2

Corresponding Author:
Alan Katz, MBChB, MSc

Abstract

Background: Lower back pain is a common weakening condition that affects a large population. It is a leading cause of disability and lost productivity, and the associated medical costs and lost wages place a substantial burden on individuals and society. Recent advances in artificial intelligence and natural language processing have opened new opportunities for the identification and management of risk factors for lower back pain. In this paper, we propose and train a deep learning model on a data set of clinical notes that have been annotated with relevant risk factors, and we evaluate the model's performance in identifying risk factors in new clinical notes.

Objective: The primary objective is to develop a novel deep learning approach to detect risk factors for underlying disease in patients presenting with lower back pain in clinical encounter notes. The secondary objective is to propose solutions to potential challenges of using deep learning and natural language processing techniques for identifying risk factors in electronic medical record free text and make practical recommendations for future research in this area.

Methods: We manually annotated clinical notes for the presence of six risk factors for severe underlying disease in patients presenting with lower back pain. Data were highly imbalanced, with only 12% (n=296) of the annotated notes having at least one risk factor. To address imbalanced data, a combination of semantic textual similarity and regular expressions was used to further capture notes for annotation. Further analysis was conducted to study the impact of downsampling, binary formulation of multi-label classification, and unsupervised pretraining on classification performance.

Results: Of 2749 labeled clinical notes, 347 exhibited at least one risk factor, while 2402 exhibited none. The initial analysis shows that downsampling the training set to equalize the ratio of clinical notes with and without risk factors improved the macro–area under the receiver operating characteristic curve (AUROC) by 2%. The Bidirectional Encoder Representations from Transformers (BERT) model improved the macro-AUROC by 15% over the traditional machine learning baseline. In experiment 2, the proposed BERT–convolutional neural network (CNN) model for longer texts improved (4% macro-AUROC) over the BERT baseline, and the multitask models are more stable for minority classes. In experiment 3, domain adaptation of BERTCNN using masked language modeling improved the macro-AUROC by 2%.

Conclusions: Primary care clinical notes are likely to require manipulation to perform meaningful free-text analysis. The application of BERT models for multi-label classification on downsampled annotated clinical notes is useful in detecting risk factors suggesting an indication for imaging for patients with lower back pain.

(JMIR Med Inform 2023;11:e45105) doi:10.2196/45105

KEYWORDS

machine learning; lower back pain; natural language processing; semantic textual similarity; electronic medical records; risk factors; deep learning

Introduction

Lower back pain (LBP) is recognized as a common disability worldwide [1-3]. While there is no agreed-upon definition of LBP, in a systematic review, it was primarily defined through routinely collected electronic health data, which include *International Classification of Diseases, Ninth Revision* (*ICD-9*) and *International Statistical Classification of Diseases, Tenth Revision* (*ICD-10*) codes [4]. One estimate of the burden of LBP is that 13% of adults in the United States live with LBP, while in Canada, among those living with chronic pain, 50.9% identified the location of their pain in the upper or lower back

[2,3]. In a systematic review [4], the mean prevalence of LBP among the studies collected ranged between 1.4% and 15.6%.

While the burden of LBP remains high, it is important to understand the indicators for possible serious underlying causes that require imaging, also known as "risk factors" [5]. According to Choosing Wisely Canada, risk factors may include [6]:

- A history of cancer
- Unexplained weight loss
- A recent infection
- Fever
- Loss of bowel or bladder control
- Abnormal reflexes or the loss of muscle power in the legs

Radiological (diagnostic) imaging includes procedures such as x-rays, computed tomography scans, or magnetic resonance imaging scans. Recommendations from clinical practice guidelines state that, unless risk factors are present, radiological imaging is not needed for patients with LBP [5,7]. Moreover, ordering radiological imaging when it is unnecessary puts the patient at risk for radiation exposure and other negative consequences [5,6]. Despite these recommendations, patients with LBP are frequently subjected to unnecessary imaging [8].

The data for this study in clinical practice uses electronic medical records (EMRs). The widespread use of this IT has introduced the feasibility of analyzing large numbers of clinical notes without having to manually access paper charts and perform the analyses using automated approaches such as natural language processing (NLP) [9]. The Canadian Primary Care Sentinel Surveillance Network [10] routinely extracts clinical information such as clinical encounter notes, note type, and the date of the notes from primary care clinical practices with the permission of the providers. Applying NLP methods to EMR data makes it possible to detect LBP risk factors and understand the use of imaging in this common clinical presentation.

Since the introduction of transformers in 2019 [11], which are large language models that can be fine-tuned for specific tasks, deep language models have achieved a significant milestone in natural language understanding. The transfer learning paradigm of unsupervised pretraining and fine-tuning [12] using Bidirectional Encoder Representations from Transformers (BERT) has reduced the requirement for large labeled data sets to achieve state-of-the-art analytic performance. Previous research [13] has explored the use of topic models and deep neural networks to automatically distinguish acute LBP episodes using free-text clinical notes.

Methods

The following steps were undertaken to achieve our goal: preparation of EMR data, EMR annotation process, addressing imbalanced data, and application of the proposed model.

Preparation of EMR Data

We accessed a random sample of deidentified EMR data, and using the regular expressions created in SAS (SAS Institute), we identified a cohort of patients with any indication of LBP. Notes were further filtered by note type to only include provider-generated clinical notes. The data were then split randomly into three files. Ethics approval for the study was provided by the University of Manitoba Health Research Ethics Board and the Health Information Privacy Committee.

EMR Annotation Process

Six medical students reviewed the EMR notes to identify the six LBP risk factors in accordance to Choosing Wisely Canada. They worked in teams of two to validate the application of the inclusion and exclusion criteria, each note being annotated by two students. The inclusion criteria listed in Textbox 1 were the presence of specific clinical notes suggestive of at least one of the six risk factors indicating the need for imaging. The exclusion criteria were the presence of clinical conditions that could lead to symptoms that may be confused with any of the underlying conditions represented by the six risk factors and clinical notes that do not represent relevant visits.

Textbox 1. Inclusion and exclusion criteria for risk factors.

Inclusion criteria

- Lower extremities for loss of muscle function
- Positive straight leg test
- Nerve impingement
- Sciatica, but need to confirm radiculopathy
- Incontinence related to a nerve issue
- If back pain has improved
- Follow-up discussions of imaging results
- Saddle anesthesia
- Notes that do not specify upper vs lower back pain

Exclusion criteria

- HIV is not a relevant infection (regardless of viral load and strain/location)
- Urinary symptoms other than incontinence are neither risk factors nor symptoms of relevant infection
- Shingles as an infection if it is a lumbar dermatome
- Nocturnal enuresis
- Degenerative diseases or osteoarthritis with an indication of back pain
- Copy/pasted imaging results onto the electronic medical record note
- Notes that mention previous or resolved back pain
- Well child/adolescent visit

An experienced clinician (AK) arbitrated any disagreements between student annotators. This supported the inclusion of correctly labeled records in the classification model. For the annotation process, we used Microsoft Forms (Microsoft Corporation), which enabled us to collect the relevant data in a systematic and organized manner. Specifically, the output from Microsoft Forms was linked to a secure CSV file containing the clinical notes, using a unique identifier to facilitate data merging and subsequent analysis.

Addressing Imbalanced Data

Our data collection process consisted of two rounds. In the first round, we established the initial distribution of risk factors. Analysis of this round revealed an imbalanced distribution of labels, a well-known factor that can impact the performance of deep learning methods [14,15]. Specifically, we observed an imbalance in both the infrequent occurrence of individual risk factors and the high frequency of the "null class," which denotes the absence of risk factors.

To address this imbalance, we adopted a 2-pronged approach. First, we collected additional clinical notes specifically targeting minority risk factors. Second, we downsampled the majority of notes with "null class." Notably, the initial data set lacked any clinical notes for unexplained weight loss. Table 1 depicts the distribution of risk factors after the first labeling round, revealing that only 12% (n=296) of the 2487 annotated notes exhibited any risk factors.

Table . Risk factor distribution after the first labeling round. Zero notes exhibit the unexplained weight loss risk factor.

Risk factors	Annotations (round 1), n
Cancer	26
Weight[a]	0
Fever	8
Infection	8
Bowel	9
Abreflex	233

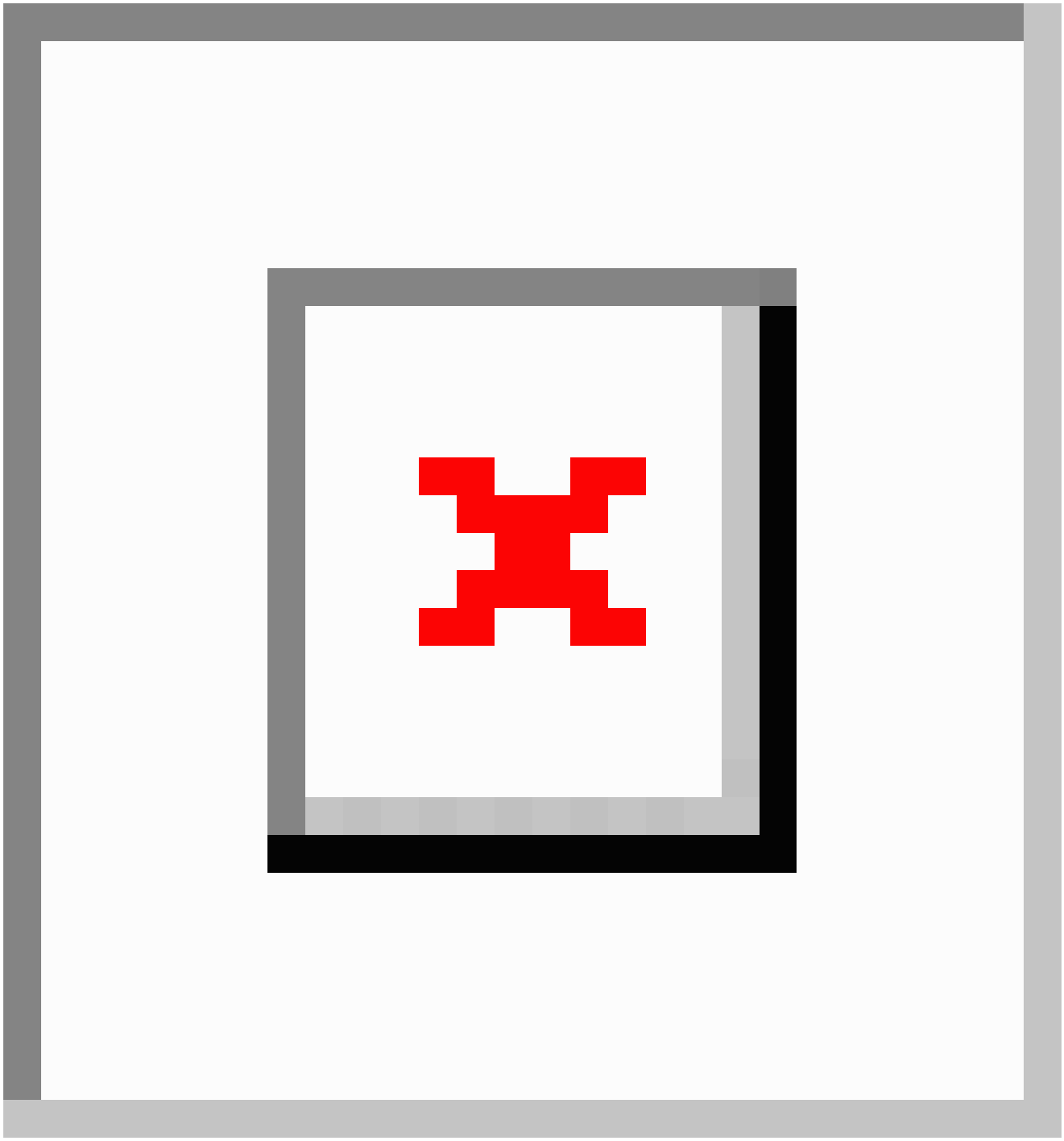
[a]Zero notes exhibit the unexplained weight loss risk factor.

Acquiring More Notes to Annotate

Prior studies have explored methods for addressing the challenge of obtaining sufficient data for training [16]. To acquire clinical notes for labeling that are more likely to exhibit a minority risk factor, we used unsupervised semantic textual similarity (STS). It is a ranking task where given a text query and a list of clinical notes, the STS model ranks the clinical notes that are semantically like the query. We trained two unsupervised STS models, Transformers and Sequential Denoising Auto-Encoder (TSDAE) [17] and Simple Contrastive Learning of Sentence Embeddings (SimCSE) [18], implemented using the SentenceTransformer Python library [19]. To rank the unlabeled clinical notes (ie, 55,000 notes with any LBP indication), we formed the queries using rationales, collected as part of the first labeling round. Here, we refer to "rationale" as an extracted snippet or text from the clinical note the annotators highlighted as evidence for a risk factor.

Figure 1 illustrates the STS sampling process with numbered steps. First, we group the clinical notes based on the exhibited risk factors. We then concatenate the rationales for each group of clinical notes to form queries and rank the unlabeled clinical notes using the unsupervised STS models. If the rationales were unavailable from the first labeling round (eg, "weight loss"), we used risk factor definition or custom text as the query. We selected the top K notes from the ranked clinical notes, where "K" is set within the 10-50 range. We further filtered noisy outputs using phrases such as "has fever," "has back pain," and "lost weight." Finally, we iterated the process for each risk factor and provided the selected notes for the second labeling round.

Figure 1. Semantic textual similarity sampling process, followed for the second labeling round. STS: semantic textual similarity.



This approach helped maximize annotations for clinical notes that exhibited risk factors. Table 2 depicts the complete distribution of risk factors after both rounds of labeling. Of the 262 annotated clinical notes in the second round, 19.5% (n=51) of the clinical notes exhibited risk factors, in contrast to 12% (n=296) in the first round.

Table . Risk factor distribution after both rounds of labeling.[a]

Risk factors	Annotations (round 1 + 2), n
Cancer	53
Weight	32
Fever	17
Infection	9
Bowel	9
Abreflex	236

[a]This includes 2487 notes from the first round and 262 notes from the second round. In the second labeling round, we collected 32 clinical notes for the unexplained weight loss risk factor.

Treating Class Imbalance With Downsampling

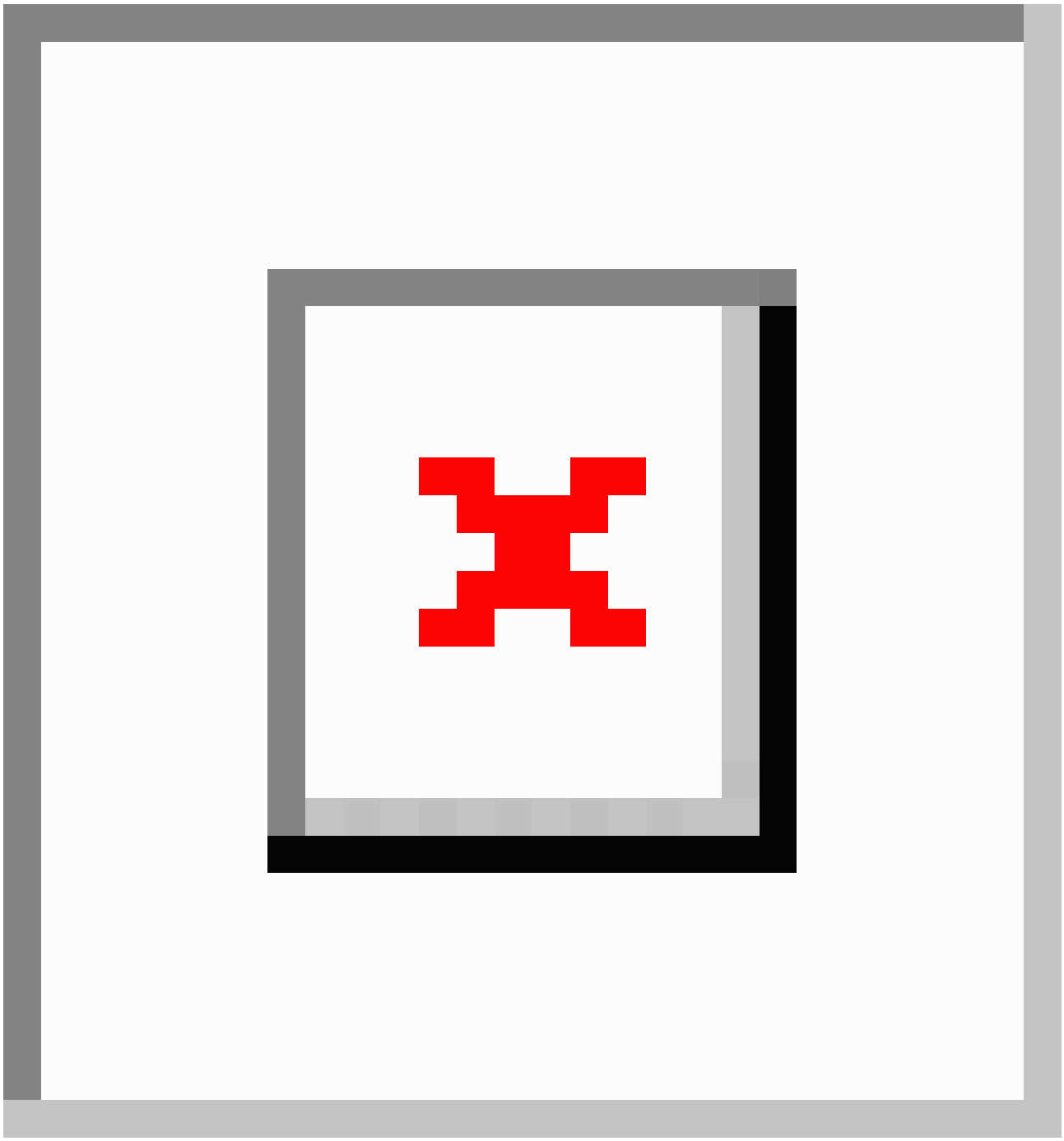
Following the second round of labeling, a significant class imbalance was observed in the resulting distribution of labels. Specifically, out of the total 2749 annotated clinical notes, only 347 were labeled as having one or more risk factors, while the remaining 2402 notes were labeled with no risk factor. To mitigate this issue, two common approaches are oversampling the minority class or downsampling the majority class. In a multi-label data set, each instance can be assigned to one or more classes. For instance, in the case of clinical notes, they may have one or more risk factors, making it challenging to oversample the minority class. This is because generating synthetic instances requires randomly selecting a minority clinical note that may have a combination of labels rather than a single label. However, this approach may bias the model toward the minority class and lead to overfitting. Consequently, we opted for downsampling the majority class to balance the class distribution and prevent the model from being biased toward the majority class.

Specifically, a subset of the clinical notes with "no risk factors" was randomly selected to match the number of clinical notes with "any risk factor." This approach aimed to balance the class distribution and enable the model to learn from both positive and negative examples. To assess the effectiveness of the downsampling strategy, we conducted a comparative analysis of the model's performance with and without downsampling.

Application of Proposed Model

Transformer-based BERT [11] models can be fine-tuned for detecting risk factors in clinical notes using a small labeled data set. The requirement for large labeled data sets is eased with models that are pretrained on large clinical text. In this work, we used BlueBERT [20] as our back-end model that is pretrained on PubMed abstracts and clinical notes from the Medical Information Mart for Intensive Care (MIMIC-III) data set [21]. However, BERT models are limited to a maximum input length of 512 tokens. The length of clinical notes in our data set ranges from 7 to 1400 tokens with 8% (n=221) of the notes having more than 512 tokens. To overcome this limitation, we propose a novel architecture called BERT–convolutional neural network (CNN) that chunks the inputs and processes them using convolution layers. The proposed chunking method is illustrated in Figure 2.

Figure 2. BERT input chunking: a clinical note is first separated into chunks of 512 tokens. Each chunk is then independently processed by the BERT-based back-end model. The chunk embedding is obtained by averaging the token embeddings from the last five layers of BERT. Finally, all the chunk embeddings are concatenated and processed using convolution layers, as defined by Kim [22]. Note: The sample clinical note does not belong to the real data set. BERT: Bidirectional Encoder Representations from Transformers.



Experimental Setup

The study used a repeated 2-fold cross-validation approach with two repetitions to improve the estimated performance of the machine learning models. As the data set was multi-label, we adopted the iterative stratification method [23,24] provided by the scikit-multilearn library [25] to generate stratified splits for the folds. This ensured that every split had a similar distribution of risk factors. The 2-fold cross-validation was repeated twice, resulting in a total of four runs. Wherever applicable, we implemented the downsampling technique (as described earlier) on the training set. Our results are reported in terms of the area under the receiver operating characteristic curve (AUROC) of individual risk factors and their macroaverage across the folds. Table 3 reports the frequency of positive risk factors in each split of the folds.

Table . Frequency of positive risk factors in train-test splits. We report the approximate counts of each risk factor across folds. Note: the counts do not include the clinical notes with no risk factors, which are approximately 1198 and 1195 for the train and test split, respectively.

Positive risk factors	Train split (n=1374 notes), n	Test split (n=1375 notes), n
Cancer	26	27
Weight	16	16
Fever	8	9
Infection	4	5
Bowel	4	5
Abreflex	118	118

Ethics Approval

The study received ethics approval from the Health Research Ethics Board of the University of Manitoba (study number HS20263; review number H2016:408).

Results

Overview

In this section, we report the analysis of the data collection and classification performance of the transformer-based models with different configurations, including traditional machine learning and BERT-based baselines. The transformer-based models were trained for 10 epochs each, with a learning rate ranging from 5e-05 to 6e-5. Unless specified otherwise, all the BERT-based models use BlueBERT [20] as the back end.

Data Collection Analysis

Each annotation was added to the clinical note level independently. These notes are associated with patient- and site-level information, allowing for further analysis based on the patient and site as the unit of analysis. Table 4 presents an analysis of the LBP characteristics reported in the collected data, using notes, patient, and site ID as the units of analysis. This enables a multilevel analysis of the reported characteristics, providing a detailed understanding of their distribution across various units of analysis.

Table . Lower back pain characteristics gathered from collected data, with notes, patient, and site ID each serving as the units of analysis.

Unit of analysis		Values, n (%)
Notes (N=2749)		
	History of cancer	53 (1.9)
	Signs of fever	17 (0.6)
	Unexplained weight loss	32 (1.2)
	Recent infection	9 (0.3)
	Loss of bowel or bladder control	9 (0.3)
	Abnormal reflexes	236 (8.6)
Patients (N=1943)		
	History of cancer	40 (2.1)
	Signs of fever	17 (0.9)
	Unexplained weight loss	32 (1.6)
	Recent infection	9 (0.5)
	Loss of bowel or bladder control	8 (0.4)
	Abnormal reflexes	201 (10.3)
Site ID (N=22)		
	History of cancer	12 (55)
	Signs of fever	11 (50)
	Unexplained weight loss	12 (55)
	Recent infection	5 (23)
	Loss of bowel or bladder control	7 (32)
	Abnormal reflexes	13 (59)

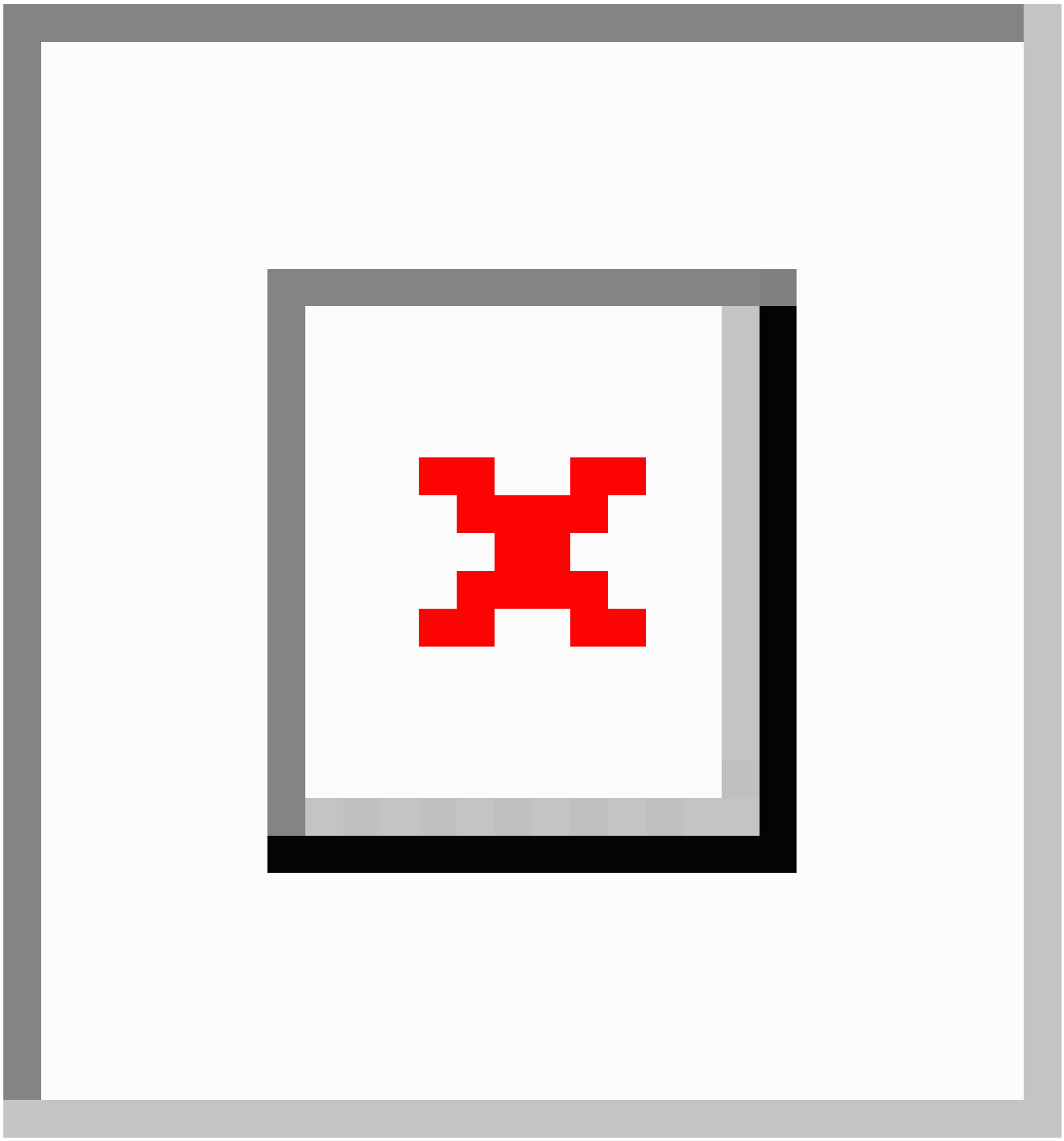
A total of 2749 clinical notes were annotated to collect information on risk factors for LBP. The most reported risk factor was "abnormal reflexes," with 236 annotations, followed by "history of cancer" with 53 annotations. Out of the 1943 patients covered by the annotation process, only 40 were labeled with a "history of cancer," accounting for 2.1% (n=40) of the total patients. More than 10% of patients were reported with "abnormal reflexes," while "recent infection" and "loss of bowel control" were reported in only 9 and 8 patients, respectively.

The analysis of clinical sites associated with the clinical notes revealed that 12 of 22 sites reported at least two risk factors, with "recent infection" and "loss of bowel or bladder control" being the least commonly reported risk factors, mentioned in only 5 and 7 clinical sites, respectively. These findings indicate that "abnormal reflexes" is the most reported characteristic of LBP across all units of analysis, with "history of cancer," "unexplained weight loss," and "signs of fever" being reported less frequently. The frequency of "loss of bowel or bladder control" and "recent infection" was relatively low across all units of analysis, indicating that these characteristics may not be as common as others in cases of LBP. The distribution of these characteristics varies across different units of analysis, which highlights the importance of examining LBP characteristics at multiple levels.

Performance With and Without Downsampling

In our initial analysis, we compared the impact of downsampling the training set, as described earlier, on the average and label-wise performance of the models. Figure 3 displays the results of this comparison. We also included a tf-idf (term frequency–inverse document frequency) + logistic regression model trained with a multi-output classifier [26] as a baseline, which was the best-performing baseline (among 7 candidates, including k-nearest neighbors, naive Bayes, random forest, and models from the scikit-multilearn Library [25]). On average, the BERT models performed 15% better than the baseline. Downsampling the training set improved performance by 2% for BERT-Multi models and reduced the SD as reflected by the error bars for minority labels (eg, "bowel" and "fever"). Downsampling of the majority class (ie, "No Risk factor notes") also helped stabilize the performance of the models, as indicated by the smaller error bars. We used the downsampled training set for further analysis.

Figure 3. Comparison of BERT-Multitask models trained on complete and downsampled data. A tf-idf + logistic regression model trained with a multi-output classifier is included as a baseline. The AUROC for each risk factor and their macroaverage are reported, with the SDs reflected in the error bars. AUROC: area under the receiver operating characteristic curve; BERT: Bidirectional Encoder Representations from Transformers; tf-idf: term frequency–inverse document frequency.

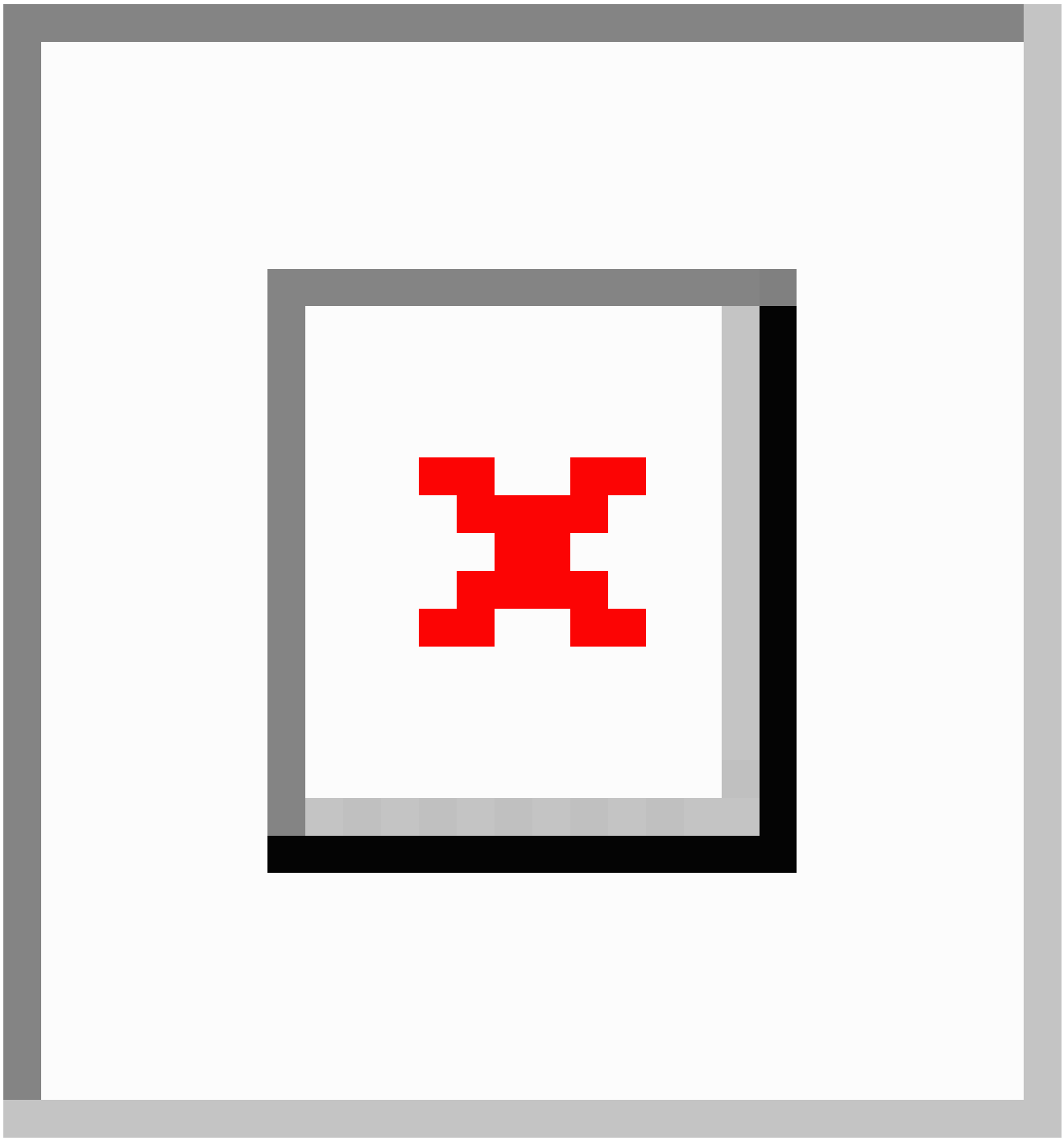


Performance With BERTCNN and Independent Binary Classifiers

Using the downsampled training set for all the models, we compared the performance of four different models chosen by architecture (BERT, BERTCNN) and task formulation (multitask learning, binary classification). Figure 4 shows the results. The comparison of BERT and BERTCNN highlights the importance of not truncating longer inputs. The comparison of the proposed model (BERTCNN) with their binary variants helps in understanding the trade-off between parameter efficiency and performance. The average AUROC of all the models are comparable, with BERTCNN-Multi performing 4% better than BERT-Multi. The multitask BERT and BERTCNN models match the performance of their binary alternative with six times fewer parameters. When sufficient positive samples are present for a risk factor (eg, abreflex), all the models perform comparably with a low SD. When the samples are insufficient (eg, “infection” and “bowel”), the binary models have high SD (indicated by the error bars), as few-samples BERT fine-tuning is known to be unstable [27]. In such cases, the multitask models generally produce more stable results, with the BERTCNN-Multi performing 9% better than BERT-Multi. In general, the BERTCNN model can benefit from the extra context

found in the complete clinical note to improve prediction performance.

Figure 4. BERT-Multi, BERT-Binary, BERTCNN-Multi, and BERTCNN-Binary trained on the downsampled training data. The AUROC for each risk factor and their macroaverage are reported, with the SDs reflected in the error bars. AUROC: area under the receiver operating characteristic curve; BERT: Bidirectional Encoder Representations from Transformers; BERTCNN: Bidirectional Encoder Representations from Transformers–convolutional neural network.

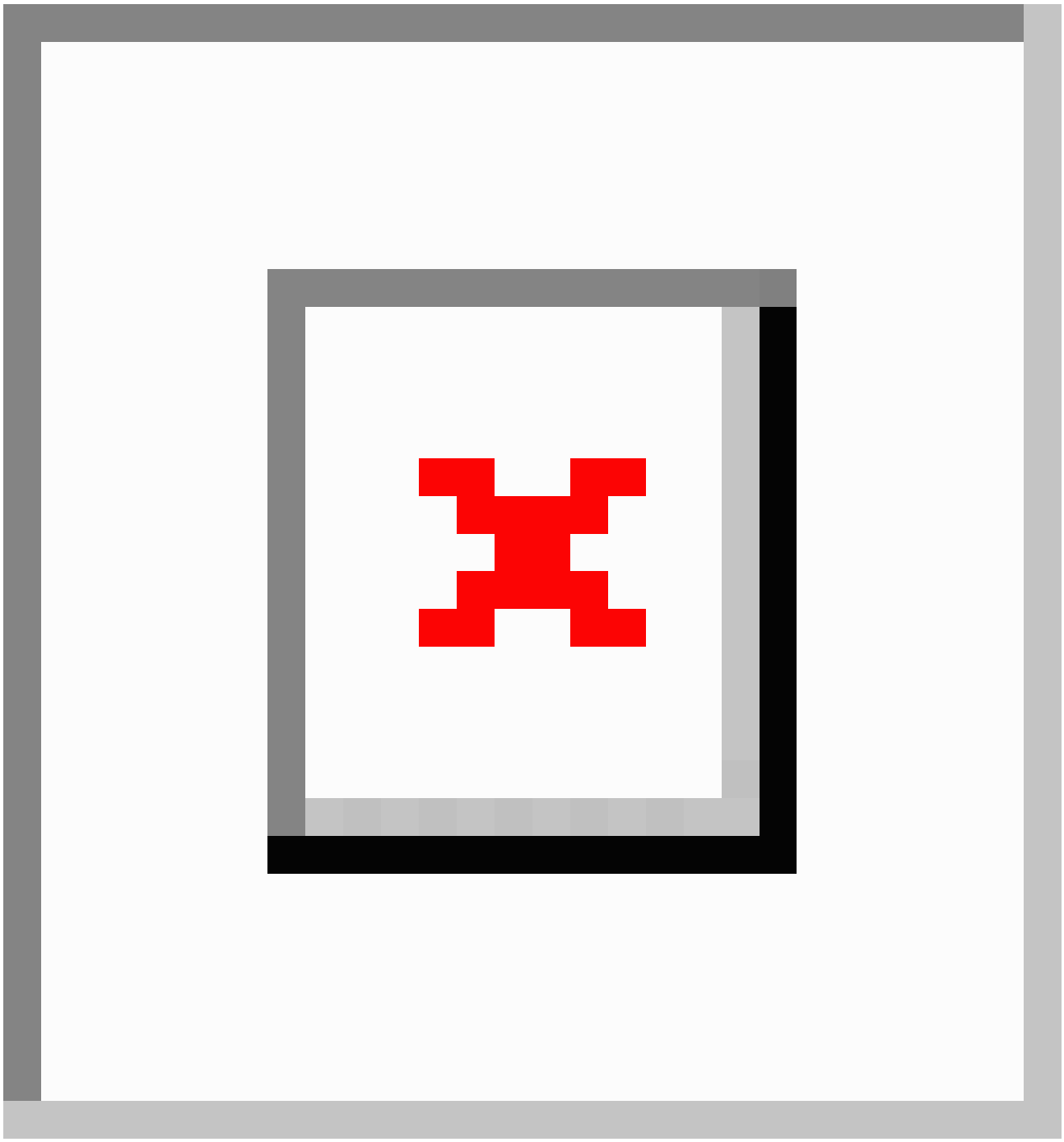


Performance With Domain Adaptation Using Unsupervised Training

The best-performing model can further benefit from pretraining [28] the underlying transformer model using the clinical notes. In this analysis, we investigate the effect of domain adaptation using pretraining on classification performance. We used BERTCNN and further pretrained the back-end model (BlueBERT [20]) with the complete corpus of relevant clinical notes (N=57,000) for 3 epochs. Two choices for pretraining the BERT architecture were considered: masked language modeling (MLM; BERTCNN-MLM-Multi) [12] and causal language modeling (CLM; BERTCNN-CLM-Multi) [29]. In addition, we also report results of the recent transformers-based model for long text in the clinical domain, called clinical-longformer [30,31], which was pretrained on clinical notes from the MIMIC-III data set [21]. Our results, shown in Figure 5, indicate that the MLM method performed 2% better than no domain adaptation and improved the performance for “cancer” by 5%. The longformer model further improves performance over MLM

by 2%. It is worth noting that while the performance improvement of domain adaptation using MLM [32] is not significant, it is comparable to that of the already pretrained BlueBERT [20] and clinical-longformer [30,31], which were pretrained on a much larger corpus of over 2 million notes.

Figure 5. Effect of domain adaptation using MLM, CLM, and comparison with the clinical-longformer model. The AUROC for each risk factor and their macroaverage are reported, with the SDs reflected in the error bars. AUROC: area under the receiver operating characteristic curve; BERTCNN: Bidirectional Encoder Representations from Transformers–convolutional neural network; CLM: causal language modeling; MLM: masked language modeling.



Discussion

The analysis of electronic clinical notes using machine learning techniques provides the opportunity to explore and evaluate clinical care, previously not possible when clinical experts had to read each clinical record. NLP of clinical records is still a relatively new research endeavor that is rapidly evolving. This study encountered and addressed several challenges that are likely to be common in the analysis of clinical notes. For example, the initially collected data were imbalanced, with most clinical notes having no risk factors for cancer, suggesting the need for further investigation of LBP. By sampling additional clinical notes from the unlabeled pool using unsupervised semantic matching techniques for a limited second round of labeling, we captured 7.5% more clinical notes with at least one risk factor. Strategic resampling can decrease bias in multi-label

data sets, which substantially helps in classification performance. The analysis comparing multitask learning and binary classification suggests we can match the performance of independent binary classifiers and produce more stable results while using a fraction of the learned parameters required for binary classifiers. This study demonstrates the value of domain adaptation as an additional technique to improve the classification results of transformer-based models and improve clinical free-text classification using unsupervised methods.

A strength of this study is the comparison of different models and approaches using a random sample of real clinical notes. We compared the BERT-based model, which does not truncate longer clinical notes and uses the complete context to make predictions, to the more commonly used truncated note model. The extensive empirical analysis on the impact of different modeling choices, including comparisons of multitask and single-task learning, resampling of data, and domain adaptation using unsupervised methods for the detection of LBP risk factors in clinical notes, provides guidance for future analysis of clinical text data.

While the low number of samples for certain risk factors in the test set is a limitation, this was addressed in reporting the AUROC for each individual risk factor, including their macroaverage for each model, and using the repeated k-fold cross-validation approach for better estimation of performance.

Future research will involve linking the outcomes of imaging studies to the identification of risk factors in this data set. It is anticipated that patients without risk factors would have normal imaging, while those with risk factors should be more likely to have abnormal imaging suggestive of disease requiring further treatment. Those analyses will need to address the imbalance in the data, as a minority of patients have undergone imaging.

Deep learning models, specifically BERT-based models, are suitable for capturing and detecting risk factors for LBP in clinical notes. Semantic matching techniques are effective during data collection in providing minority samples for labeling and improving data set distribution. The proposed method BERTCNN can be successfully applied for clinical notes that may be longer than the input limit of BERT-based models. Detecting risk factors in clinical notes is better formulated as multitask learning, which is more efficient and provides stable results. Furthermore, transformer-based models are successfully adopted for clinical text using transfer learning and MLM.

Acknowledgments

The authors acknowledge the clinicians and patients whose data were accessed for this study through the Manitoba Primary Care Research Network, a node of the Canadian Sentinel Surveillance Network. The authors also thank medical student annotators Elvina Mukhamedshina, Gem Newman, JaeYeon Park, Mehrin Ahmed, Sue Zhang, and Will Siemens.

Conflicts of Interest

None declared.

References

1. Centers for Disease Control and Prevention. Acute low back pain. 2022. URL: web.archive.org/web/20220709154456/www.cdc.gov/acute-pain/low-back-pain/index.html [accessed 2022-06-1]
2. Stevans JM, Delitto A, Khoja SS, Patterson CG, Smith CN, Schneider MJ, et al. Risk factors associated with transition from acute to chronic low back pain in US patients seeking primary care. JAMA Netw Open 2021 Feb 1;4(2):e2037371. [doi: 10.1001/jamanetworkopen.2020.37371] [Medline: 33591367]
3. MacDougall HL, George SZ, Dover GC. Low back pain treatment by athletic trainers and athletic therapists: BIOMEDICAL or Biopsychosocial orientation? J Athl Train 2019 Aug 6;54(7):772-779. [doi: 10.4085/1062-6050-430-17] [Medline: 31386578]
4. Fatoye F, Gebrye T, Odeyemi I. Real-world incidence and prevalence of low back pain using routinely collected data. Rheumatol Int 2019 Mar 8;39(4):619-626. [doi: 10.1007/s00296-019-04273-0] [Medline: 30848349]
5. Chou R. Low back pain. Ann Intern Med 2021 Aug 10;174(8):ITC113-ITC128. [doi: 10.7326/AITC202108170] [Medline: 34370518]
6. Choosing Wisely Canada. Imaging tests for lower back pain. 2022. URL: choosingwiselycanada.org/pamphlet/imaging-tests-for-lower-back-pain/ [accessed 2022-06-2]
7. Bach SM, Holten KB. Guideline update: what's the best approach to acute low back pain? J Fam Pract 2009;58(12):E1. [Medline: 19961812]]
8. Rao D, Scuderi G, Scuderi C, Grewal R, Sandhu SJ. The use of imaging in management of patients with low back pain. J Clin Imaging Sci 2018 Aug 24;8:30. [doi: 10.4103/jcis.JCIS_16_18] [Medline: 30197821]
9. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in Healthcare. Nat Med 2019 Jan;25(1):24-29. [doi: 10.1038/s41591-018-0316-z] [Medline: 30617335]
10. Birtwhistle RV. Canadian Primary Care Sentinel Surveillance Network: a developing resource for family medicine and public health. Can Fam Physician 2011 Oct;57:10-1221. [Medline: 21998241]]

11. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. 2017 Presented at: Presented at Advances in Neural Information Processing Systems 30 (NIPS 2017); December 4-9; Long Beach, CA URL: papers.nips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html
12. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep Bidirectional transformers for language understanding. 2019 Presented at: Presented at Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers); June; Minneapolis, MN p. 4171-4186. [doi: 10.18653/v1/N19-1423]
13. Miotto R, Percha BL, Glicksberg BS, Lee HC, Cruz L, Dudley JT, et al. Identifying acute low back pain episodes in primary care practice from clinical notes: observational study. JMIR Med Inform 2020 Feb 27;8(2):e16878. [doi: 10.2196/16878] [Medline: 32130159]
14. Japkowicz N, Stephen S. The class imbalance problem: a systematic study. Intelligent Data Analysis 2002;6(5):429-449. [doi: 10.3233/IDA-2002-6504]
15. Krawczyk B. Learning from imbalanced data: open challenges and future directions. Prog Artif Intelligence 2016 Nov;5(4):221-232. [doi: 10.1007/s13748-016-0094-0]
16. Humbert-Droz M, Mukherjee P, Gevaert O. Strategies to address the lack of labeled data for supervised machine learning training with electronic health records: case study for the extraction of symptoms from clinical notes. JMIR Med Inform 2022 Mar 14;10(3):e32903. [doi: 10.2196/32903] [Medline: 35285805]
17. Wang K, Reimers N, Gurevych I. TSDAE: using transformer-based sequential denoising auto-encoder for unsupervised sentence embedding learning. 2021 Presented at: Presented at Findings of the Association for Computational Linguistics: EMNLP 2021; November; Punta Cana, Dominican Republic p. 671-688. [doi: 10.18653/v1/2021.findings-emnlp.59]
18. Gao T, Yao X, Chen D. Simcse: simple Contrastive learning of sentence Embeddings. 2021 Presented at: Presented at Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing; November; Online and Punta Cana, Dominican Republic p. 6894-6910. [doi: 10.18653/v1/2021.emnlp-main.552]
19. Reimers N, Gurevych I. Sentence-BERT: sentence embeddings using siamese BERT-networks. 2019 Presented at: Presented at Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); November; Hong Kong, China p. 3982-3992. [doi: 10.18653/v1/D19-1410]
20. Peng Y, Yan S, Lu Z. Transfer learning in BIOMEDICAL natural language processing: an evaluation of BERT and Elmo on ten benchmarking datasets. 2019 Presented at: Presented at Proceedings of the 18th BioNLP Workshop and Shared Task; August; Florence, Italy p. 58-65. [doi: 10.18653/v1/W19-5006]
21. Johnson AEW, Pollard TJ, Shen L, Lehman LWH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016 May 24;3:160035. [doi: 10.1038/sdata.2016.35] [Medline: 27219127]
22. Kim Y. Convolutional neural networks for sentence classification. 2014 Presented at: Presented at Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); October; Doha, Qatar p. 1746-1751. [doi: 10.3115/v1/D14-1181]
23. Szymański P, Kajdanowicz T. A network perspective on stratification of multi-label data. 2017 Presented at: Presented at Proceedings of the First International Workshop on Learning With Imbalanced Domains: Theory and Applications; September 22; Skopje, Macedonia p. 22-35 URL: proceedings.mlr.press/v74/szyma%C5%84ski17a.html
24. Sechidis K, Tsoumakas G, Vlahavas I. On the stratification of multi-label data. In: Gunopulos D, Hofmann T, Malerba D, Vazirgiannis M, editors. Machine Learning and Knowledge Discovery in Databases, Part III: European Conference, ECML PKDD 2010, Athens, Greece, September 5-9, 2011, Proceedings, Part III. Berlin, Heidelberg: Springer; 2011:145-158. [doi: 10.1007/978-3-642-23808-6]
25. Szymański P, Kajdanowicz T. A Scikit-based python environment for performing multi-label classification. arXiv. Preprint posted online on February 5, 2017 . [doi: 10.48550/arXiv.1702.01460]
26. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2011;12(85):2825-2830.
27. Zhang T, Wu F, Katiyar A, Weinberger KQ, Artzi Y. Revisiting few-sample BERT fine-tuning. 2021 Presented at: Presented at 9th International Conference on Learning Representations; May 3-7; Virtual Event, Austria URL: openreview.net/forum?id=cO1IH43yUF
28. Gururangan S, Marasović A, Swayamdipta S, Lo K, Beltagy I, Downey D, et al. Don't stop pretraining: adapt language models to domains and tasks. 2020 Presented at: Presented at Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; Online p. 8342-8360. [doi: 10.18653/v1/2020.acl-main.740]
29. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov R, Le QV. XLNet: generalized Autoregressive Pretraining for language understanding. arXiv. Preprint posted online on June 19, 2019 . [doi: 10.48550/arXiv.1906.08237]
30. Li Y, Wehbe RM, Ahmad FS, Wang H, Luo Y. Clinical-Longformer and clinical-Bigbird: transformers for long clinical sequences. arXiv. Preprint posted online on January 27, 2022 . [doi: 10.48550/arXiv.2201.11838]
31. Li Y, Wehbe RM, Ahmad FS, Wang H, Luo Y. A comparative study of pretrained language models for long clinical text. J Am Med Inform Assoc 2023 Jan 18;30(2):340-347. [doi: 10.1093/jamia/ocac225] [Medline: 36451266]

32. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. 2019 Presented at: Presented at Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers); June; Minneapolis, MN p. 4171-4186. [doi: 10.18653/v1/N19-1423]

Abbreviations

AUROC: area under the receiver operating characteristic curve
BERT: Bidirectional Encoder Representations from Transformers
BERTCNN: Bidirectional Encoder Representations from Transformers–convolutional neural network
CLM: causal language modeling
EMR: electronic medical record
***ICD-10*:** *International Statistical Classification of Diseases, Tenth Revision*
***ICD-9*:** *International Classification of Diseases, Ninth Revision*
LBP: lower back pain
MIMIC-III: Medical Information Mart for Intensive Care
MLM: masked language modeling
NLP: natural language processing
SimCSE: Simple Contrastive Learning of Sentence Embeddings
STS: semantic textual similarity
tf-idf: term frequency–inverse document frequency
TSDAE: Transformers and Sequential Denoising Auto-Encoder

Edited by C Lovis; submitted 15.12.22; peer-reviewed by G Liu, H Wang, R Abeysinghe; revised version received 11.05.23; accepted 03.06.23; published 09.08.23.

Please cite as:
Jaiswal A, Katz A, Nesca M, Milios E
Identifying Risk Factors Associated With Lower Back Pain in Electronic Medical Record Free Text: Deep Learning Approach Using Clinical Note Annotations
JMIR Med Inform 2023;11:e45105
URL: https://medinform.jmir.org/2023/1/e45105
doi:10.2196/45105

Original Paper

Near Real-time Natural Language Processing for the Extraction of Abdominal Aortic Aneurysm Diagnoses From Radiology Reports: Algorithm Development and Validation Study

Simon Gaviria-Valencia[1], MD; Sean P Murphy[2], BS; Vinod C Kaggal[3], MS; Robert D McBane II[4], MD; Thom W Rooke[4], MD; Rajeev Chaudhry[5], MBBS, MPH; Mateo Alzate-Aguirre[1], MD; Adelaide M Arruda-Olson[1], MD, PhD

[1]Divisions of Preventive Cardiology and Cardiovascular Ultrasound, Department of Cardiovascular Medicine, Mayo Clinic, Rochester, MN, United States

[2]Advanced Analytics Services Unit (Natural Language Processing), Department of Information Technology, Mayo Clinic, Rochester, MN, United States

[3]Enterprise Technology Services (Natural Language Processing), Department of Information Technology, Mayo Clinic, Rochester, MN, United States

[4]Gonda Vascular Center, Department of Cardiovascular Medicine, Mayo Clinic, Rochester, MN, United States

[5]Department of Internal Medicine, Mayo Clinic, Rochester, MN, United States

Corresponding Author:
Adelaide M Arruda-Olson, MD, PhD
Divisions of Preventive Cardiology and Cardiovascular Ultrasound
Department of Cardiovascular Medicine
Mayo Clinic
200 First Street SW
Rochester, MN, 55905
United States
Phone: 1 507 266 6853
Email: arrudaolson.adelaide@mayo.edu

Abstract

Background: Management of abdominal aortic aneurysms (AAAs) requires serial imaging surveillance to evaluate the aneurysm dimension. Natural language processing (NLP) has been previously developed to retrospectively identify patients with AAA from electronic health records (EHRs). However, there are no reported studies that use NLP to identify patients with AAA in near real-time from radiology reports.

Objective: This study aims to develop and validate a rule-based NLP algorithm for near real-time automatic extraction of AAA diagnosis from radiology reports for case identification.

Methods: The AAA-NLP algorithm was developed and deployed to an EHR big data infrastructure for near real-time processing of radiology reports from May 1, 2019, to September 2020. NLP extracted named entities for AAA case identification and classified subjects as cases and controls. The reference standard to assess algorithm performance was a manual review of processed radiology reports by trained physicians following standardized criteria. Reviewers were blinded to the diagnosis of each subject. The AAA-NLP algorithm was refined in 3 successive iterations. For each iteration, the AAA-NLP algorithm was modified based on performance compared to the reference standard.

Results: A total of 360 reports were reviewed, of which 120 radiology reports were randomly selected for each iteration. At each iteration, the AAA-NLP algorithm performance improved. The algorithm identified AAA cases in near real-time with high positive predictive value (0.98), sensitivity (0.95), specificity (0.98), F1 score (0.97), and accuracy (0.97).

Conclusions: Implementation of NLP for accurate identification of AAA cases from radiology reports with high performance in near real time is feasible. This NLP technique will support automated input for patient care and clinical decision support tools for the management of patients with AAA.

(JMIR Med Inform 2023;11:e40964) doi:10.2196/40964

KEYWORDS

abdominal aortic aneurysm; algorithm; big data; electronic health record; medical records; natural language processing; radiology reports; radiology

Introduction

Worldwide prevalence rates of abdominal aortic aneurysms (AAAs) range from 1.6% to 3.3% for men older than 60 years [1]. Assessment of AAA may be performed by a variety of imaging tests, including ultrasound (US), computerized tomography (CT), and magnetic resonance imaging (MRI). In the United States, the prevalence of AAA has been reported as 2.8% among 9457 individuals screened by US [2]. Moreover, screening for early identification decreases the risk of aneurysm-related death and morbidity [1,3]. A prior study has shown that 4.5 ruptured AAA per 10,000 person-years were likely to have been prevented by screening, with an estimated 54 life-years gained per year of screening in a population of 23,000 men at risk [4].

The interpretation of imaging examinations is routinely reported in radiology reports as narrative text in electronic health records (EHRs) [5]. The automated extraction of information from narrative text can be accomplished by natural language processing (NLP) [6-8]. Prior studies have demonstrated high accuracy, sensitivity, specificity, and positive predictive value (PPV) of NLP for extraction of clinical concepts from narrative text in radiology reports [9-12]. Moreover, NLP is useful in cohort ascertainment for epidemiologic studies, query-based case retrieval, clinical decision support (CDS), quality assessment of radiologic practices, and diagnostic surveillance [5].

A previous retrospective cohort study from our institution developed a rule-based NLP algorithm for retrospective retrieval of AAA cases from radiology reports, which performed with high accuracy [12]. However, to the best of our knowledge, no prior study has demonstrated the use of NLP to identify AAA cases from radiology reports processed in near real-time. Hence, we tested the hypothesis that a rule-based NLP algorithm will extract AAA diagnosis from radiology reports in near real-time with high accuracy.

Methods

Study Settings

This study used Mayo Clinic radiology reports from May 1, 2019, to September 30, 2020.

Study Design

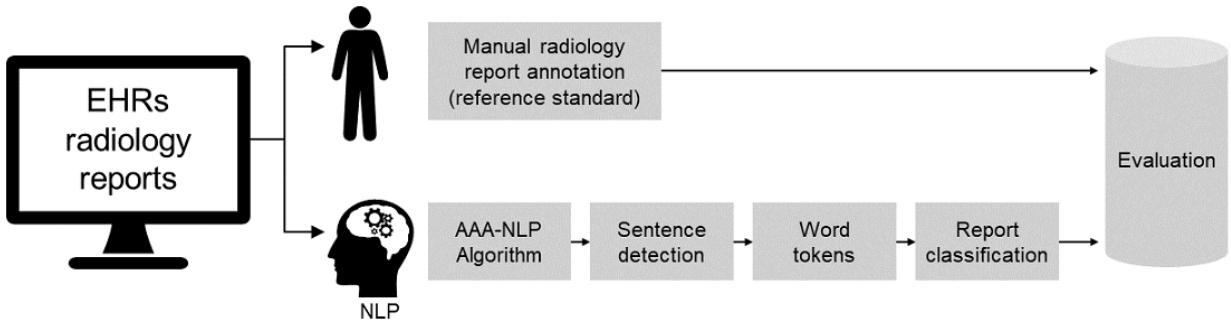
A rule-based AAA-NLP algorithm was developed for information extraction of AAA diagnosis automatically from radiology reports, including CT abdomen pelvis without intravenous (IV) contrast, CT chest abdomen pelvis angiogram with IV contrast, US abdomen complete, US aorta iliac arteries bilateral with doppler, MRI abdomen with and without IV contrast, and MRI pelvis with and without IV contrast. The rule-based NLP algorithm was developed using MedTagger and deployed in the institutional near real-time big data infrastructure to process relevant radiology reports. MedTagger is an open-source NLP tool that has been previously used in various clinical NLP applications [13]. MedTagger enables section identification, extraction of concepts, sentences, and word tokenization [14,15]. The AAA-NLP algorithm had 2 main components composed of text processing and report classification. AAA-relevant concepts were used to classify all reports (Figure 1).

A custom lexicon for AAA was identified by the study team through a manual review of radiology reports. Subsequently, this lexicon was mapped to corresponding concepts and their synonyms in the Unified Medical Language System Metathesaurus. The lexicon used for AAA identification included aorta abdominal aneurysm, aortic aneurysm abdominal, AAA, aneurysm abdominal aorta, and infrarenal aortic aneurysm. Each radiology report was then processed in near real-time by NLP. The AAA-NLP algorithm extracted both the lexicon and the contextual information of assertions, including negations or confirmations, from each radiology report. Textbox 1 displays the rules used by the NLP algorithm. The AAA-NLP algorithm classified subjects as AAA cases and controls without AAA.

To enable validation, the NLP output generated by near real-time processing of radiology reports was retrieved from the digital infrastructure by the information technology team and converted to a human-readable format for annotation. This annotation was performed by 2 trained physicians following written guidelines for standardization. The annotators were blinded to the diagnosis of each subject and to the results of the other annotator. In the written guidelines, AAA was defined as an aortic aneurysm diameter ≥3 cm by imaging as recommended by clinical practice guidelines [16].

The annotators reviewed the output from 120 processed radiology reports in 3 different training sets for iterative validation cycles to refine the algorithm. A total of 360 reports were reviewed. After abstracting and classifying the radiology reports, the information was entered and stored in a digital data set. Reports with a diagnosis of AAA were categorized as “case”; if there was no evidence of AAA or if an alternate diagnosis other than AAA was reported, the report was categorized as “control.” A board-certified cardiologist verified the information and resolved discrepancies in patient classification.

Figure 1. Study design. AAA: abdominal aortic aneurysm; EHR: electronic health record; NLP: natural language processing.



Textbox 1. Abdominal aortic aneurysm (AAA)–natural language processing rule and examples of text span.

Rule (any token + keyword for AAA + any token)

Examples of confirmatory assertions

- Suprarenal *aortic abdominal aneurysm* which measures up to 5.2 cm
- Fusiform *infrarenal abdominal aortic aneurysm* terminating proximal to the aortobiiliac bifurcation, 56 mm, previously 56 mm
- There is a 5.7×5.1 cm *infrarenal aortic aneurysm* measured on image 175 of series 4

Examples of negated assertions

- *Negative* for *abdominal aortic aneurysm* or dissection
- *Abdominal aortic aneurysm* is *absent*
- *Negative* for thoracic or *abdominal aortic aneurysm*, dissection, penetrating atherosclerotic ulcer or intramural hematoma

Statistical Analysis

The information extracted by the AAA-NLP algorithm from radiology reports in near real-time was compared to the reference standard manual review of radiology reports following written guidelines for standardization to calculate PPV, sensitivity, specificity, and F1 score. The formula to calculate F1 score was given as follows: 2 × ((PPV×sensitivity) / (PPV+sensitivity)) [5].

Ethics Approval

This project was approved by the Mayo Clinic Institutional Review Board (approval number 21-006950).

Results

Reports of 295 patients were validated in 3 different iterations. The data set for each iteration contained 120 reports, but 46 (16%) patients had more than one report. The reasons for more than one report for the same patient were imaging tests performed before and after repair procedures or surveillance for serial assessment of AAA (Table 1). There were no discrepancies regarding AAA diagnosis between 2 or more imaging reports from the same patient. Table 1 shows the distribution of demographic characteristics across AAA cases and controls. Cases and controls had similar ages in each of the iterative validation cycles, and most patients were Caucasian. AAA cases were more likely to have a history of smoking.

For evaluation of the AAA-NLP algorithm performance, 120 processed reports from each iteration were randomly selected. A total of 360 processed reports were reviewed by 2 physicians blinded to AAA diagnosis. There was 100% agreement for interactions 1 and 3. For interaction 2, the annotators disagreed on 1 report yielding a kappa coefficient of 92%. The disagreement was resolved by a board-certified cardiologist, creating the reference standard for comparison. The number of reports classified by the reference standard as true positives, false positives, true negatives, and false negatives in each iteration is shown in Table 2.

Table 1. Clinical characteristics and radiology report information.

Characteristic	Iteration 1		Iteration 2		Iteration 3	
	Case (n=31)	Control (n=52)	Case (n=44)	Control (n=59)	Case (n=59)	Control (n=50)
Age (years), mean (SD)	78.6 (11.1)	74.4 (12.4)	70.3 (8.4)	69.5 (14.1)	81.2 (8.5)	72.8 (10.4)
Male sex, n (%)	26 (84)	21 (40)	34 (77)	33 (56)	46 (78)	25 (50)
Caucasian, n (%)	31 (100)	52 (100)	42 (95)	54 (92)	58 (98)	48 (96)
Comorbidities, n (%)						
Hypertension	24 (77)	39 (75)	31 (70)	27 (46)	46 (78)	37 (74)
Hyperlipidemia	21 (68)	22 (42)	29 (66)	29 (49)	42 (71)	27 (54)
Smoking history	29 (94)	24 (46)	35 (80)	23 (39)	50 (85)	25 (50)
DM[a]	9 (29)	7 (13)	10 (23)	12 (20)	19 (32)	13 (26)
PAD[b]	4 (13)	4 (8)	5 (11)	4 (7)	9 (15)	4 (8)
CAD[c]	16 (52)	7 (13)	18 (41)	10 (17)	32 (54)	15 (30)
Radiology reports						
Patients with ≥2 reports, n	18	7	13	1	3	4
AAA[d] diameter (cm), mean (SD)	4.6 (1.08)	N/A[e]	4.8 (1.3)	N/A	4.9 (1.2)	N/A
Reports after AAA repair, n	2	N/A	9	N/A	8	N/A

[a]DM: diabetes mellitus.

[b]PAD: peripheral artery disease.

[c]CAD: coronary artery disease.

[d]AAA: abdominal aortic aneurysm.

[e]N/A: not applicable.

Table 2. Classification of abdominal aortic aneurysm from radiology reports during iterative validation.

	Iteration 1			Iteration 2			Iteration 3		
	Predicted case	Predicted control	Total	Predicted case	Predicted control	Total	Predicted case	Predicted control	Total
Actual case	TP[a] 59	FN[b] 6	65	TP 56	FN 2	58	TP 59	FN 3	62
Actual control	FP[c] 1	TN[d] 54	55	FP 4	TN 58	62	FP 1	TN 57	58
Total	60	60	120	60	60	120	60	60	120

[a]TP: true positive.

[b]FN: false negative.

[c]FP: false positive.

[d]TN: true negative.

Radiology reports are composed of multiple sections. Figure 2 shows an example of a deidentified radiology report with all sections.

During the first iteration implementation, section ID number was used and section detection was challenging. For the second iteration, the algorithm was revised to include section header names for the filter criteria and solve sentence boundary issues. For the third iteration, section detection was implemented based on section names from our complete corpus using the frequency of normalized text with the tool lexical variant generation of the National Library of Medicine [17]. In a separate experiment, 203 additional radiology reports were reviewed by the annotators for evaluation of report section extraction, which resulted in accuracy of 0.96.

During this iterative refinement process, the report sections termed “reason for exam,” “referral diagnosis,” “exam type,” and “signed by” (Figure 2) were excluded, resulting in enhanced NLP algorithm performance. The report sections selected for processing were findings and impressions. During each iteration, the algorithm performance further improved. The performance metrics of the iterations are summarized in Table 3.

During the last iteration, 3 false negatives and 1 false positive contributed to the error analysis. False negatives were due to the complex nature of narrative text in these reports (ie, *no* significant interval changes in appearances of a partially

thrombosed infrarenal AAA measuring 42×40 mm, extending to the level of aortic bifurcation and proximal common iliac arteries; *no* signs of rupture or impending rupture of the known infrarenal AAA; and *no* slightly increased size of fusiform infrarenal AAA). Additionally, the false positive was due to a typographical error, which was the report of a patient with an aorta diameter of 2.7 cm labeled as AAA, which does not meet the criteria for AAA (≥3.0 cm).

Figure 2. Example of deidentified radiology report with all sections. In this figure, section names are displayed in blue font. AAA: abdominal aortic aneurysm.

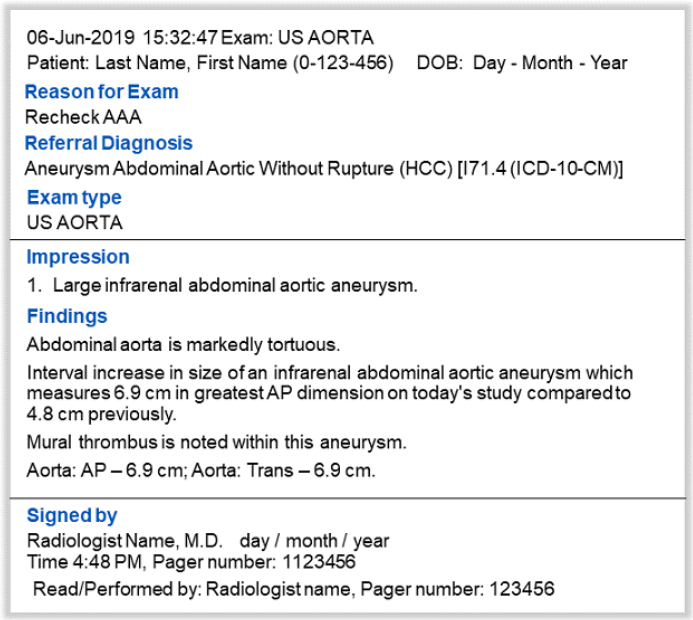


Table 3. Algorithm performance of each iteration.

Performance metric	Iteration 1 (n=120)	Iteration 2 (n=120)	Iteration 3 (n=120)
Sensitivity	0.91	0.97	0.95
PPV[a]	0.98	0.93	0.98
Specificity	0.98	0.94	0.98
F1 score	0.94	0.95	0.97
Accuracy	0.94	0.95	0.97

[a]PPV: positive predictive value.

Discussion

Overview

In this study, a novel rule-based NLP algorithm was developed for the extraction of AAA diagnosis from radiology reports and prospectively deployed in the institutional big data infrastructure for near real-time processing. Compared to the reference standard of manual review of radiology reports, the AAA-NLP algorithm extracted AAA diagnosis in near real time with high sensitivity, PPV, F1 score, specificity, and accuracy.

To the best of our knowledge, this study is the first to describe the use of NLP algorithms prospectively to extract AAA diagnosis in near real time from radiology reports. Clinicians, information technologists, and informaticians collaborated to refine the algorithm to improve performance. In previous studies, billing codes were used to find AAA cases [18,19]. However, in those studies, the cohorts were limited to patients with AAA who underwent procedures for aneurysm repair or had a history of ruptured AAA [18,19]. No prior studies using billing codes algorithms retrieved a broader spectrum of AAA diagnosis while also including patients presenting with uncomplicated AAA (ie, patients who did not undergo prior repair or who had not previously presented with ruptured AAA). In contrast, in this study, NLP automatically extracted AAA diagnosis from radiology reports prospectively and regardless of prior repair or rupture, thereby expanding the scope of computational approaches to include the detection of AAA cases prior to rupture or repair.

A radiology report consists of free text, organized into standard sections [5]. The American College of Radiology has published guidelines with recommendations for the use of sections for narrative (free text) entry in radiology reports [20]. NLP techniques enable the automatic extraction of information from narrative text [6-8]. Moreover, information extracted by NLP can be used to populate CDS systems automatically without the need for manual data entry and be better aligned with existing workflows such that radiologists can spend time interpreting images rather than filling out forms.

NLP is a computational methodology used for electronic phenotyping to extract meaningful clinical information from text fields [6,7,21]. In this study, we used NLP to process radiology text reports. The previous NLP algorithm used to find cases of AAA from radiology reports [12] was designed for retrospective cohort identification, whereas this report describes the prospective implementation of an NLP algorithm for input to a patient-specific CDS system for near real-time processing of radiology reports. Near real-time processing requires <3 milliseconds to process a document after a radiologist releases a report to the EHR [22]. The AAA-NLP implementation described in this study was developed within the existing digital infrastructure and can be used in clinical practice immediately without the need to retrain the algorithm. Additionally, the previously described algorithm [12] did not identify document sections in the radiology reports. By selecting specific sections for NLP information extraction, improvement in NLP performance was observed, as shown in the Results section. In the future, transformer-based NLP models [23,24] may be trained to interpret nuanced language, and ablation experiments [25] could be used to further evaluate these models.

The use of NLP algorithms has advantages compared to other methods. In comparison, the use of check box forms in radiology reports may require the development of new workflows [26,27]. The use of check box forms also requires the radiologist to direct attention away from the imaging interpretation process [26,27]. Manual entry of summaries of radiology findings in a check box can increase reporting time with decreased radiologist productivity [26,27]. Check box use could also result in the loss of important and clinically relevant descriptive information available only in the radiology narrative reports.

The rule-based AAA-NLP algorithm described in this study shows accurate detection of a broad spectrum of AAA cases prospectively in near real time from radiology reports, regardless of the presence of prior rupture or repair. This methodology will also potentially generate input for CDS to assist providers in managing patients with AAA by displaying the relevant information automatically at the point of care and in near real time for CDS tools. It will also support the automatic identification of cohorts for research purposes (eg, cohorts for clinical trials) and quality projects, and will support a learning health care system. NLP has been previously used for the identification of peripheral arterial disease and critical limb ischemia from narrative clinical notes of EHRs [21,28]. Therefore, it will also be possible to develop NLP algorithms for the identification of AAA cases from clinical notes in near real time.

In efforts to develop a learning health care system, Mayo Clinic has developed a robust big data–empowered clinical NLP infrastructure that enables near real-time NLP processing for the delivery of relevant information to the point of care via CDS [22]. Accordingly, we have deployed the AAA-NLP algorithm described herein to this digital infrastructure for translation to clinical practice. Importantly, the near real-time identification of patients with AAA by NLP responds to the American Heart Association scientific statement, which recommends the implementation of technologies to extract clinical information in real time that will promptly provide synopses of the information extracted [29].

Limitations

This NLP algorithm was developed, tested, and implemented in a single tertiary medical center. Future studies should evaluate this algorithm at other institutions to demonstrate portability. A robust institutional digital infrastructure is required for the execution of near real-time processing of radiology reports [22]. Hence, the absence of adequate digital infrastructure may limit porting of this algorithm. For implementation, the analysis of radiology report architecture to enable the selection of document types and document sections may also be necessary for portability. Another potential challenge for porting this algorithm to other EHRs is differences in lexicons used for the extraction of the AAA concept across institutions. In mitigation, for this NLP algorithm, each lexicon was mapped to corresponding concepts and synonyms in the publicly available Unified Medical Language System Metathesaurus for standardization.

The algorithm was developed for the extraction of AAA diagnosis but not for the extraction of iliac artery or thoracic aortic aneurysms. Future studies should create and validate NLP algorithms for the extraction of thoracic and iliac artery aneurysms. The clinical criteria for AAA diagnosis involve a minimum diameter, but this NLP algorithm did not interpret the reported diameter. This is an area for future improvement in the algorithm, as clinical criteria for AAA may change over time. In this study, most patients were Caucasian. This was likely related to the ethnic distribution of communities in the Midwest, where this study was conducted [30,31]. Additionally, prior studies have reported a higher prevalence of AAA among Caucasians compared to other races [31,32]. There were differences in comorbidities of patients included in the 3 iterations. However, the NLP was developed for the extraction of the diagnosis of AAA and not developed for the extraction of associated patient comorbidities. The differences in patient comorbidities did not influence NLP performance for the extraction of AAA from radiology reports.

Conclusions

Implementation of NLP for prospective identification of AAA cases from radiology reports in near real time with high performance is feasible. This near real-time NLP technique described will potentially be helpful for the generation of automated input for CDS tools to assist clinicians in the management of patients with AAA, quality improvement projects, and research (automated identification of cohorts).

Acknowledgments

The authors would like to thank Kara M Firzlaff for secretarial support and Christopher G Scott for statistical analysis. This study was funded by the Mayo Clinic K2R award.

Data Availability

The data sets generated and analyzed during this study are not publicly available because participants in this study did not agree for their data to be shared publicly but are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

References

1. US Preventive Services Task Force, Owens DK, Davidson KW, Krist AH, Barry MJ, Cabana M, et al. Screening for abdominal aortic aneurysm: US preventive services task force recommendation statement. JAMA 2019;322(22):2211-2218. [doi: 10.1001/jama.2019.18928] [Medline: 31821437]
2. Summers KL, Kerut EK, Sheahan CM, Sheahan MG. Evaluating the prevalence of abdominal aortic aneurysms in the United States through a national screening database. J Vasc Surg 2021;73(1):61-68. [doi: 10.1016/j.jvs.2020.03.046] [Medline: 32330595]
3. Chaikof EL, Dalman RL, Eskandari MK, Jackson BM, Lee WA, Mansour MA, et al. The society for vascular surgery practice guidelines on the care of patients with an abdominal aortic aneurysm. J Vasc Surg 2018;67(1):2-77.e2 [FREE Full text] [doi: 10.1016/j.jvs.2017.10.044] [Medline: 29268916]
4. Wilmink ABM, Quick CRG, Hubbard CS, Day NE. Effectiveness and cost of screening for abdominal aortic aneurysm: results of a population screening program. J Vasc Surg 2003;38(1):72-77 [FREE Full text] [doi: 10.1016/s0741-5214(03)00135-6] [Medline: 12844092]
5. Pons E, Braun LMM, Hunink MGM, Kors JA. Natural language processing in radiology: a systematic review. Radiology 2016;279(2):329-343. [doi: 10.1148/radiol.16142770] [Medline: 27089187]
6. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet 2012;13(6):395-405. [doi: 10.1038/nrg3208] [Medline: 22549152]
7. Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? J Biomed Inform 2009;42(5):760-772 [FREE Full text] [doi: 10.1016/j.jbi.2009.08.007] [Medline: 19683066]
8. Lopez-Jimenez F, Attia Z, Arruda-Olson AM, Carter R, Chareonthaitawee P, Jouni H, et al. Artificial intelligence in cardiology: present and future. Mayo Clin Proc 2020;95(5):1015-1039. [doi: 10.1016/j.mayocp.2020.01.038] [Medline: 32370835]
9. Wang Y, Mehrabi S, Sohn S, Atkinson EJ, Amin S, Liu H. Natural language processing of radiology reports for identification of skeletal site-specific fractures. BMC Med Inform Decis Mak 2019;19(suppl 3):73 [FREE Full text] [doi: 10.1186/s12911-019-0780-5] [Medline: 30943952]
10. Fiszman M, Chapman WW, Aronsky D, Evans RS, Haug PJ. Automatic detection of acute bacterial pneumonia from chest X-ray reports. J Am Med Inform Assoc 2000;7(6):593-604 [FREE Full text] [doi: 10.1136/jamia.2000.0070593] [Medline: 11062233]
11. Solti I, Cooke CR, Xia F, Wurfel MM. Automated classification of radiology reports for acute lung injury: comparison of keyword and machine learning based natural language processing approaches. Proceedings (IEEE Int Conf Bioinformatics Biomed) 2009;2009:314-319 [FREE Full text] [doi: 10.1109/BIBMW.2009.5332081] [Medline: 21152268]
12. Sohn S, Ye Z, Liu H, Chute CG, Kullo IJ. Identifying abdominal aortic aneurysm cases and controls using natural language processing of radiology reports. AMIA Jt Summits Transl Sci Proc 2013;2013:249-253 [FREE Full text] [Medline: 24303276]
13. Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, et al. Clinical information extraction applications: a literature review. J Biomed Inform 2018;77:34-49 [FREE Full text] [doi: 10.1016/j.jbi.2017.11.011] [Medline: 29162496]
14. Liu H, Bielinski SJ, Sohn S, Murphy S, Wagholikar KB, Jonnalagadda SR, et al. An information extraction framework for cohort identification using electronic health records. AMIA Jt Summits Transl Sci Proc 2013;2013:149-153 [FREE Full text] [Medline: 24303255]
15. Wagholikar K, Torii M, Jonnalagadda S, Liu H. Feasibility of pooling annotated corpora for clinical concept extraction. AMIA Jt Summits Transl Sci Proc 2012;2012:38 [FREE Full text] [Medline: 22779047]
16. Hirsch AT, Haskal ZJ, Hertzer NR, Bakal CW, Creager MA, Halperin JL, American Association for Vascular Surgery/Society for Vascular Surgery, Society for Cardiovascular Angiography and Interventions, Society for Vascular Medicine and Biology, Society of Interventional Radiology, ACC/AHA Task Force on Practice Guidelines. ACC/AHA guidelines for the management of patients with peripheral arterial disease (lower extremity, renal, mesenteric, and abdominal aortic): a collaborative report from the American Associations for Vascular Surgery/Society for Vascular Surgery, Society for Cardiovascular Angiography and Interventions, Society for Vascular Medicine and Biology, Society of Interventional Radiology, and the ACC/AHA Task Force on practice guidelines (writing committee to develop guidelines for the management of patients with peripheral arterial disease)--summary of recommendations. J Vasc Interv Radiol 2006;17(9):1383-1397. [doi: 10.1097/01.RVI.0000240426.53079.46] [Medline: 16990459]
17. Lexical Tools: Lvg (Lexical Variants Generation). National Library of Medicine. 2022. URL: https://tinyurl.com/2ej3v4p [accessed 2023-01-24]

18. Borthwick KM, Smelser DT, Bock JA, Elmore JR, Ryer EJ, Ye Z, et al. ePhenotyping for abdominal aortic aneurysm in the Electronic Medical Records and Genomics (eMERGE) network: algorithm development and konstanz information miner workflow. Int J Biomed Data Min 2015;4(1):113 [FREE Full text] [Medline: 27054044]
19. Jetty P, van Walraven C. Coding accuracy of abdominal aortic aneurysm repair procedures in administrative databases: a note of caution. J Eval Clin Pract 2011;17(1):91-96. [doi: 10.1111/j.1365-2753.2010.01373.x] [Medline: 20846277]
20. ACR practice parameter for communication of diagnostic imaging findings. American College of Radiology. 2020. URL: https://www.acr.org/-/media/acr/files/practice-parameters/communicationdiag.pdf [accessed 2023-01-24]
21. Afzal N, Mallipeddi VP, Sohn S, Liu H, Chaudhry R, Scott CG, et al. Natural language processing of clinical notes for identification of critical limb ischemia. Int J Med Inform 2018;111:83-89 [FREE Full text] [doi: 10.1016/j.ijmedinf.2017.12.024] [Medline: 29425639]
22. Kaggal VC, Elayavilli RK, Mehrabi S, Pankratz JJ, Sohn S, Wang Y, et al. Toward a learning health-care system: knowledge delivery at the point of care empowered by big data and NLP. Biomed Inform Insights 2016;8(suppl 1):13-22 [FREE Full text] [doi: 10.4137/BII.S37977] [Medline: 27385912]
23. Gillioz A, Casas J, Mugellini E, Khaled OA. Overview of the transformer-based models for NLP tasks. 2020 Presented at: 2020 15th Conference on Computer Science and Information Systems (FedCSIS); September 6-9, 2020; Sofia, Bulgaria p. 179-183. [doi: 10.15439/2020f20]
24. Gillioz A, Casas J, Mugellini E, Khaled OA. Overview of the transformer-based models for NLP tasks. 2020 Presented at: 2020 15th Conference on Computer Science and Information Systems (FedCSIS); September 6-9, 2020; Sofia, Bulgaria p. 179-183. [doi: 10.15439/2020f20]
25. Zhang D, Yin C, Zeng J, Yuan X, Zhang P. Combining structured and unstructured data for predictive models: a deep learning approach. BMC Med Inform Decis Mak 2020;20(1):280. [doi: 10.1186/s12911-020-01297-6]
26. Weiss DL, Langlotz CP. Structured reporting: patient care enhancement or productivity nightmare? Radiology 2008;249(3):739-747. [doi: 10.1148/radiol.2493080988] [Medline: 19011178]
27. Mityul MI, Gilcrease-Garcia B, Mangano MD, Demertzis JL, Gunn AJ. Radiology reporting: current practices and an introduction to patient-centered opportunities for improvement. AJR Am J Roentgenol 2018;210(2):376-385. [doi: 10.2214/AJR.17.18721] [Medline: 29140114]
28. Afzal N, Sohn S, Abram S, Scott CG, Chaudhry R, Liu H, et al. Mining peripheral arterial disease cases from narrative clinical notes using natural language processing. J Vasc Surg 2017;65(6):1753-1761 [FREE Full text] [doi: 10.1016/j.jvs.2016.11.031] [Medline: 28189359]
29. Maddox TM, Albert NM, Borden WB, Curtis LH, Ferguson TB, Kao DP, American Heart Association Council on Quality of Care and Outcomes Research, Council on Cardiovascular Disease in the Young, Council on Clinical Cardiology, Council on Functional Genomics and Translational Biology, Stroke Council. The learning healthcare system and cardiovascular care: a scientific statement from the American Heart Association. Circulation 2017;135(14):e826-e857 [FREE Full text] [doi: 10.1161/CIR.0000000000000480] [Medline: 28254835]
30. St Sauver JL, Grossardt BR, Leibson CL, Yawn BP, Melton LJ, Rocca WA. Generalizability of epidemiological findings and public health decisions: an illustration from the Rochester Epidemiology Project. Mayo Clin Proc 2012;87(2):151-160 [FREE Full text] [doi: 10.1016/j.mayocp.2011.11.009] [Medline: 22305027]
31. Johnson G, Avery A, McDougal EG, Burnham SJ, Keagy BA. Aneurysms of the abdominal aorta. Incidence in blacks and whites in North Carolina. Arch Surg 1985;120(10):1138-1140. [doi: 10.1001/archsurg.1985.01390340036006] [Medline: 4038055]
32. Kent KC, Zwolak RM, Egorova NN, Riles TS, Manganaro A, Moskowitz AJ, et al. Analysis of risk factors for abdominal aortic aneurysm in a cohort of more than 3 million individuals. J Vasc Surg 2010;52(3):539-548 [FREE Full text] [doi: 10.1016/j.jvs.2010.05.090] [Medline: 20630687]

Abbreviations

AAA: abdominal aortic aneurysm
CDS: clinical decision support
CT: computerized tomography
EHR: electronic health record
IV: intravenous
MRI: magnetic resonance imaging
NLP: natural language processing
PPV: positive predictive value
US: ultrasound

Edited by C Lovis; submitted 11.07.22; peer-reviewed by N Afzal, C Reeder; comments to author 16.08.22; revised version received 29.12.22; accepted 19.01.23; published 24.02.23.

Please cite as:
Gaviria-Valencia S, Murphy SP, Kaggal VC, McBane II RD, Rooke TW, Chaudhry R, Alzate-Aguirre M, Arruda-Olson AM
Near Real-time Natural Language Processing for the Extraction of Abdominal Aortic Aneurysm Diagnoses From Radiology Reports: Algorithm Development and Validation Study
JMIR Med Inform 2023;11:e40964
URL: https://medinform.jmir.org/2023/1/e40964
doi:10.2196/40964
PMID:36826984

Original Paper

Deep Learning Approach for Negation and Speculation Detection for Automated Important Finding Flagging and Extraction in Radiology Report: Internal Validation and Technique Comparison Study

Kung-Hsun Weng[1], MSc, MD; Chung-Feng Liu[2], PhD; Chia-Jung Chen[3], MSc

[1]Department of Medical Imaging, Chi Mei Medical Center, Chiali, Tainan, Taiwan

[2]Department of Medical Research, Chi Mei Medical Center, Tainan, Taiwan

[3]Department of Information Systems, Chi Mei Medical Center, Tainan, Taiwan

Corresponding Author:
Chia-Jung Chen, MSc
Department of Information Systems
Chi Mei Medical Center
No.901, Zhonghua Rd.
Yongkang Dist.
Tainan, 71004
Taiwan
Phone: 886 6 2812811 ext 52069
Email: carolchen@mail.chimei.org.tw

Abstract

Background: Negation and speculation unrelated to abnormal findings can lead to false-positive alarms for automatic radiology report highlighting or flagging by laboratory information systems.

Objective: This internal validation study evaluated the performance of natural language processing methods (NegEx, NegBio, NegBERT, and transformers).

Methods: We annotated all negative and speculative statements unrelated to abnormal findings in reports. In experiment 1, we fine-tuned several transformer models (ALBERT [A Lite Bidirectional Encoder Representations from Transformers], BERT [Bidirectional Encoder Representations from Transformers], DeBERTa [Decoding-Enhanced BERT With Disentangled Attention], DistilBERT [Distilled version of BERT], ELECTRA [Efficiently Learning an Encoder That Classifies Token Replacements Accurately], ERNIE [Enhanced Representation through Knowledge Integration], RoBERTa [Robustly Optimized BERT Pretraining Approach], SpanBERT, and XLNet) and compared their performance using precision, recall, accuracy, and F_1-scores. In experiment 2, we compared the best model from experiment 1 with 3 established negation and speculation-detection algorithms (NegEx, NegBio, and NegBERT).

Results: Our study collected 6000 radiology reports from 3 branches of the Chi Mei Hospital, covering multiple imaging modalities and body parts. A total of 15.01% (105,755/704,512) of words and 39.45% (4529/11,480) of important diagnostic keywords occurred in negative or speculative statements unrelated to abnormal findings. In experiment 1, all models achieved an accuracy of >0.98 and F_1-score of >0.90 on the test data set. ALBERT exhibited the best performance (accuracy=0.991; F_1-score=0.958). In experiment 2, ALBERT outperformed the optimized NegEx, NegBio, and NegBERT methods in terms of overall performance (accuracy=0.996; F_1-score=0.991), in the prediction of whether diagnostic keywords occur in speculative statements unrelated to abnormal findings, and in the improvement of the performance of keyword extraction (accuracy=0.996; F_1-score=0.997).

Conclusions: The ALBERT deep learning method showed the best performance. Our results represent a significant advancement in the clinical applications of computer-aided notification systems.

(JMIR Med Inform 2023;11:e46348) doi:10.2196/46348

KEYWORDS

radiology report; natural language processing; negation; deep learning; transfer learning; supervised learning; validation study; Bidirectional Encoder Representations from Transformers; BERT; clinical application; radiology

Introduction

Background

Timely and effective communication of test results is essential in modern medicine. To promptly address patients' problems, hospitals must ensure that the test results are completed without delay and that clinicians are aware of substantial abnormal findings. Delayed or failed communication of important findings by the department performing the test and the clinical team can increase the risk of adverse patient events and result in medical malpractice and compensation, especially for potentially life-threatening and important diagnoses [1].

Although radiology reports are the primary method of communication between radiology and clinical departments, the fact that a radiologist produces a report does not necessarily mean that the clinician reads it entirely. Ignácio et al [2] showed that only 55.7% of clinicians read the entire report thoroughly. Reda et al [3] showed that >40% of clinicians read only the conclusions or only read the conclusions in detail. More than 30% of clinicians have made preventable medical errors because they did not read radiology reports carefully. Even if the radiologist has made the correct diagnosis in the report, the clinician may still miss it.

To address these communication issues, current radiology guidelines [4] now require radiologists to go beyond report completion and use additional communication methods for reports with significant findings, including flagging or alerting the report, e-mailing, or direct verbal communication via telephone. Natural language processing can also automatically extract data from radiology reports, for example, automatically extracting important diagnoses, follow-up data, or management recommendations or automatically identifying reports that require specific action [5]. These methods can help to identify important information in radiology reports or reports that need to be read in detail to alert clinicians.

In addition, the laboratory information system (LIS) used in hospitals today can automatically highlight abnormalities found in tests and display them differently to ensure that clinicians do not miss important findings, such as using different colors or special symbols [6]. For example, in our hospital, if a patient has undergone a routine blood test and some of the blood cell counts are abnormal, the LIS will automatically display the results on the computer screen in a unique color for the abnormal values and a typical color for the others. The LIS also displays important keywords (eg, nodules) within radiology reports in different colors.

However, because most radiology reports are freely typed by radiologists in an unstructured manner, both techniques encounter challenges. Negative and speculative statements are significant problems.

Radiologists can use negative statements to communicate the absence of specific diagnoses and provide a clearer picture of the patient's condition. For example, the statement "No definite CT evidence of aortic dissection" informs the clinician that the patient's condition is not related to aortic dissection.

The diagnoses in the speculative statements may or may not be related to the actual abnormal findings. The radiology report may contain speculative statements in the presence of an imaging finding of uncertain significance that requires further investigation, for example, "RUL lung nodule. Lung cancer should be suspected." In such cases, the diagnoses (lung cancer) in the speculative statements are related to abnormal findings. Even if the radiologist finds no problems with the study, the radiology report may still contain speculative statements to prevent potential medicolegal issues. Disclaimer (eg, "10%-15% of cases of breast cancer are missed on mammograms" [7]) or statement of limitations (eg, "non-enhanced images, small lesion may be obscured") are common examples. In such cases, the diagnoses (breast cancer or lesion) in the speculative statements are unrelated to the actual diagnoses.

A notification system that does not distinguish whether diagnostic Information is contained in negative or speculative statements unrelated to abnormal findings and annotates or extracts all of them to "alert" the clinician may generate excessive false alarms. Excessive false alarms can overload the clinician's senses and lead to the "cry wolf" phenomenon, causing alarm fatigue. Consequently, clinicians may delay detection or even ignore truly valuable alerts, posing a risk to patients, especially if the percentage of false alarms is high [8].

This study aimed to address the potential analytical inaccuracies resulting from negative and speculative statements in radiology reports and to facilitate the use of unstructured reports by hospital information systems.

Prior Work

Current studies have adopted various approaches to detect negation and speculation, including rule-based, machine learning–based, and deep learning–based approaches [9-17].

The rule-based approach relies on experts to define the rules that are understandable to humans. NegEx, proposed by Chapman et al [18]; NegFinder, proposed by Mutalik et al [19]; NegHunter, proposed by Gindl et al [20]; and NegExpander, proposed by Aronow et al [21], are regular expression-based approaches. Regular expression-based methods have limitations, such as the inability to capture the syntactic structure and the possibility of misinterpreting the scope of the negative and speculative statements. For example, "No change of tumor" may be misinterpreted as both "No change" and "No tumor."

Methods such as DEEPEN (Dependency Parser Negation), proposed by Mehrabi et al [22], and NegBio, proposed by Peng et al [23], analyze the syntactic structure based on grammar. These methods are more accurate than regular expression-based approaches in limiting the scope of negative and speculative statements and reducing false positives because these methods consider the dependency relationship between words. However,

these methods have certain limitations. For example, errors in the analysis may occur if the grammar of the text deviates from typical norms, such as the presence of long noun phrases [23]. When analyzing text, most of these methods [18-20,22,23] split the text into sentences that are analyzed independently. The algorithms and expert-defined rules only consider a single sentence at once and do not consider both the preceding and following contexts.

With the advancement of artificial intelligence, machine learning techniques have been applied to detect negation and speculation. For example, Medlock et al [24] proposed a weakly supervised learning–based approach to predict the labels of training samples for machine learning training and used the trained models to detect speculation in biomedical texts. Rokach et al [25] compared several machine learning approaches, including the Hidden Markov Model, Conditional Random Field (CRF), decision tree, and AdaBoost, cascaded decision tree classifiers with and without the Longest Common Sequence. They found that the cascaded decision tree with the Longest Common Sequence performed best. Morante et al proposed k-nearest neighbor algorithm–based [26] and meta-learning–based approaches [27]. Ou et al [28] compared rule-based and support vector machine–based machine learning methods and obtained better performance of machine learning methods.

Later studies began investigating deep learning–based approaches and achieved better results than previous non–deep learning approaches. Qian et al [17] were the first to propose a deep learning method for negation and speculation detection using a convolutional neural network–based model by using the relative position of tokens and path features from syntactic trees as features.

By contrast, recurrent neural networks and their derivatives, such as Long Short-Term Memory (LSTM), are suitable for processing sequential data. These architectures can incorporate dependencies on preceding and following elements, making them particularly useful for natural language processing tasks, and have achieved good results in recognizing negations and speculations. For example, in a study by Fancellu et al [14], a Bidirectional LSTM (BiLSTM)–based model was applied, and it demonstrated better performance than other methods on the Sherlock data set. Lazib et al [9] compared methods, including LSTM, BiLSTM, Gated Recurrent Unit, and CRF, and showed that the recurrent neural network–based architecture performed the best. Gautam et al [15] compared several LSTM-based models and obtained the best performance using 2-layer encoders and decoders with dropouts. Taylor et al [10] applied the BiLSTM-based model to the analysis of negation in electroencephalography reports. Sergeeva et al [11] proposed an LSTM-based approach and investigated the effect of expert-provided negation cues on the detection performance of the negation scopes. Sykes et al [12] compared the methods based on BiLSTM and feedforward neural networks and rule-based methods, including pyConText, NegBio, and EdIE-R, for negation detection in radiology reports. The BiLSTM-based approach outperformed other approaches.

BERT (Bidirectional Encoder Representations from Transformers) [29], proposed by Google in 2018, is a pretrained, transformer-based model that is effective for negation detection. Khandelwal et al [16] developed NegBERT and, in another study [13], used a multitasking approach with BERT, XLNet, and RoBERTa (Robustly Optimized BERT Pretraining Approach) for negation and speculation detection, with improved results on BioScope and Simon Fraser University review data sets compared with the control methods. Zavala et al [30] proposed a system based on BiLSTM with CRF and fine-tuned BERT; evaluated the methods on English and Spanish clinical, biomedical, and review text; and showed improved performance compared with previous methods. They also found that pretrained word embedding, especially contextualized embedding, helped to understand the biomedical text.

Numerous variants of BERT have been developed to improve performance and simplify the model. ALBERT (A Lite BERT) [31] reduces the model parameters and improves the performance through parameter sharing and matrix decomposition. DistilBERT (Distilled version of BERT) [32] uses knowledge distillation to reduce the size and improve the inference speed while retaining most of the language understanding. XLNet [33] implements autoregressive training while preserving the advantages of autoencoding models and outperforms BERT on 20 tasks. RoBERTa [34] improves the training method to outperform BERT and XLNet. ERNIE (Enhanced Representation through Knowledge Integration) [35] uses an alternative masking method to outperform BERT in Chinese tasks. SpanBERT [36] extends BERT with span-based masking and an additional training objective, resulting in a better performance on span-based tasks. DeBERTa (Decoding-Enhanced BERT With Disentangled Attention) [37] improves BERT and RoBERTa with decoupled attention, improved mask encoder, and virtual adversarial training and outperforms RoBERTa-Large on the Multigenre Natural Language Inference, Stanford Question Answering Data set, and Reading Comprehension data set from examinations tasks and humans on the SuperGLUE task. ELECTRA (Efficiently Learning an Encoder That Classifies Token Replacements Accurately) [38] outperforms BERT with a new pretraining task, Replaced Token Detection, and performs similarly to RoBERTa and XLNet with one-fourth the computation.

Contribution of This Work

This study has implications for optimizing the performance of hospital information systems in managing unstructured electronic medical records. The key findings and results of this study are as follows.

First, we found that fine-tuned general-purpose transformer models could outperform NegEx, NegBio, and NegBERT, which are explicitly designed for negation and speculation detection. We identified sources of error in the latter 3 methods and suggested potential improvements.

Second, we found that transformer, unlike NegEx and NegBio, demonstrated the ability to perform multisentence contextual analysis and further granular classification of speculative statements as related or unrelated to abnormal findings. This capability can improve information filtering in hospital information systems to eliminate nondiagnostically relevant information.

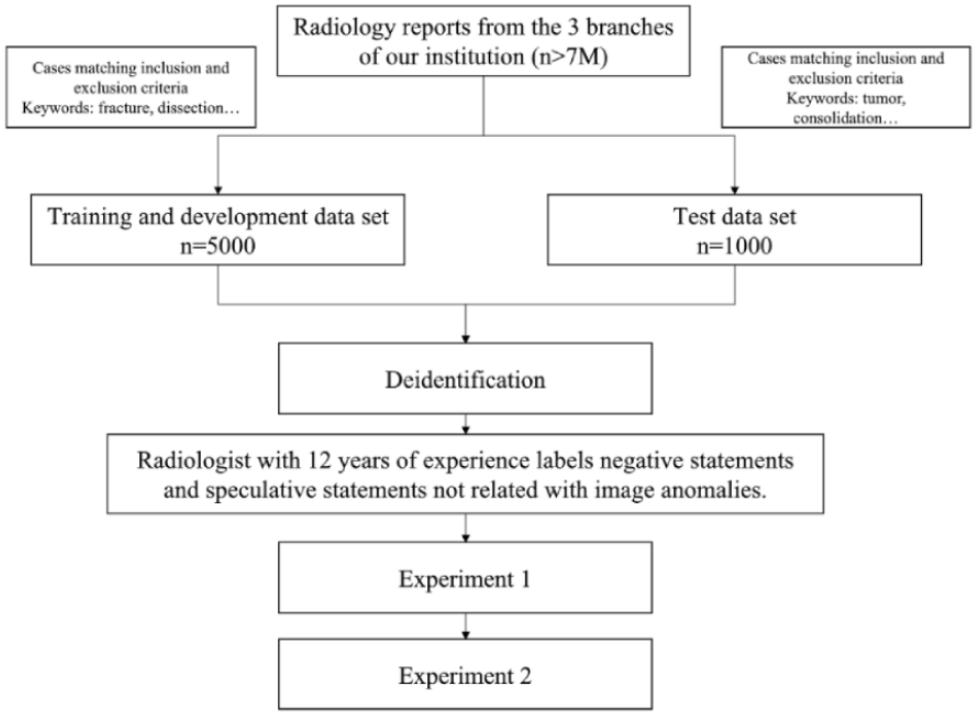
Finally, in contrast to other studies using BERT [16,39], we found that using a lightweight transformer model and learning the cues and scopes of negative and speculative sentences in a single step can perform well.

Methods

Ethics Approval

The Chi Mei Hospital Institutional Review Board reviewed and approved this study (11105-J02). This study is a retrospective analysis study using deidentified electronic medical records, thus obviating the requirement for obtaining informed consent from the individuals. Figure 1 shows the flow diagram of the study.

Figure 1. Research flow. n: number of reports.



Inclusion and Exclusion Criteria

The inclusion criteria for this study were radiological examinations performed in the 3 branches of our institution between 2012 and 2022, with the reports being written in English language and the type of examination being x-ray, special radiology, computed tomography (CT), magnetic resonance imaging (MRI), or ultrasound. We included cases that met all criteria. The exclusion criteria were Chinese reports and patients aged <20 years at the time of examination. We excluded cases that met any of the exclusion criteria. Samples were collected using 2 independent keyword searches in a search engine targeting radiology reports that met the inclusion criteria but not the exclusion criteria.

Data

Overview

The training and development data set consisted of 5000 radiology reports randomly selected from a keyword search using the terms “fracture,” “dissection,” “infarct,” “pneumothorax,” “extravasation,” “thrombosis,” or “pneumoperitoneum.” The test data set consisted of 1000 reports selected from a keyword search using the terms “tumor,” “consolidation,” “pulmonary TB,” “metastasis,” or “bleeding.” Keywords were selected from our institution’s list of important keywords and randomly assigned to the data sets. These keywords are referred to as “important keywords” in the study. The samples in the training and development and test data sets were mutually exclusive with no overlap.

The training and development data set was automatically partitioned into training and development data sets in a 9:1 ratio for model training. The training, development, and test data sets ratio was 9:1:2, with 4500, 500, and 1000 radiology reports, respectively.

In this study, each word or token was assigned to one of the 2 categories, as shown in Table 1: “Positive statements, or speculative statements potentially related to abnormal findings” (category 0) and “negative statements, or speculative statements not related to abnormal findings” (category 1). We combined speculative statements unrelated to abnormal findings with negative statements as a single class because of their limited representation. The rationale for category 1 is that the information conveyed is not relevant to abnormal findings and should not trigger highlights or alerts. A token is the minimum output unit of the transformer-based model’s tokenizer.

All radiology reports included in the study were deidentified by removing identifying information such as medical record

number, application number, examination date, ordering department, and examination time. A radiologist with 12 years of experience (KHW) reviewed the reports and annotated all negative and speculative statements unrelated to abnormal findings using the open-source Doccano [40] software. The annotation served as the gold standard for subsequent analysis.

Table 1. Classification of words and tokens in this study.

Type[a] and subtype	Example	Category[b]
Negative	Liver laceration at S6 *without active contrast extravasation*	1
Speculative		
Unrelated to abnormal findings	No CT[c] evidence of large infarct. *Suggest MRI[d] to exclude hyperacute infarct if indicated*	1
Potentially related to abnormal findings	*Rt[e] cerebellum acute infarct cannot be ruled out.*	0
Positive	*Rt cerebellum acute infarct*	0

[a]Type refers to the type of statement.

[b]Token category in the italicized text if italicization is used. All texts without italics were classified as category 0. Category 0: positive statements or speculative statements potentially related to abnormal findings. Category 1: negative statements or speculative statements not related to abnormal findings.

[c]CT: computed tomography.

[d]MRI: magnetic resonance imaging.

[e]Rt: right.

Included Negations

This study included all statements in which the radiologist explicitly denied a diagnosis or a finding. Our data included morphological negation and sentential negation, which are common forms of negative statements in English text [22]. Morphological negation involves using prefixes, such as “un-” or “ir-,” to modify certain words to express negation. Sentential negation involves using negative words, such as “no” or “without,” to negate part of the statement. In addition, radiologists at the authors’ hospital often use unique symbols or abbreviations, such as “(−)” or “[−].”

Included Speculations

In cases where the imaging study is inconclusive but there is still the possibility of a significant abnormality, the information system should notify the clinician and allow the clinician to make the final decision. Therefore, for the task of speculation detection, our focus was limited to speculative statements that were unrelated to abnormal findings. Meanwhile, we treated speculative statements that may correlate with actual abnormal findings as equivalent to positive statements.

After reviewing the samples, we identified 2 scenarios in which speculative statements could be confidently determined to be unrelated to abnormal findings. First, the radiologist explicitly stated that there was no relevant abnormality. Second, the radiologist stated that certain diagnoses could not be evaluated owing to study limitations. In all the other scenarios, speculative statements may be associated with abnormal findings.

In the following 3 examples, we classify the diagnoses or findings written in italics as speculative statements unrelated to abnormal findings. The actual test results were normal or unrelated to these diagnoses or findings.

1. No CT evidence of pulmonary embolism. Suggest V/Q scan to exclude *small branch embolism* if indicated.
2. No CT evidence of large infarct. Suggest MRI to exclude *hyperacute infarct* if indicated.
3. *Liver tumor* cannot be excluded by noncontrast CT.

In the following 2 examples, the diagnoses or findings written in italics are speculative statements considered potentially related to actual abnormal findings:

1. Equivocal filling defect in RLL segmental pulmonary artery. Suggest V/Q scan to exclude *small branch embolism* if indicated.
2. Rt cerebellum *acute infarct* cannot be ruled out.

Design of the Experiments

We conducted 2 experiments to evaluate the ability of general all-purpose pretrained deep learning models and existing negation and speculation-detection algorithms to identify negation and speculation in real-world radiology reports.

In experiment 1 (Figure 2), we fine-tuned several transformer-based models using our training and validation data sets. We performed token category prediction (category 0 or 1) for all tokens in the training, validation, and test data sets.

In experiment 2 (Figure 3), we compared 3 negation and speculation-detection algorithms that performed well on public data sets with the best model from experiment 1. The algorithms evaluated were NegEx, NegBio, which has predefined expert rules and open-source implementation, and NegBERT, whose training code is available. We then performed category prediction (category 0 or 1) for all words that matched a given “important keyword” in the test data set. We also analyzed the sources of errors. In addition, we compared the performance of keyword extraction in positive and speculative statements potentially related to abnormal findings before and after applying various algorithms.

Figure 2. Experiment 1. X: the original text, ŷ: class predicted by the model; y: the gold standard. Category 0: positive statements or speculative statements potentially related to abnormal findings; category 1: negative statements or speculative statements unrelated to abnormal findings. ALBERT: A Lite Bidirectional Encoder Representations From Transformers; BERT: Bidirectional Encoder Representations From Transformers; DeBERTa: Decoding-Enhanced Bidirectional Encoder Representations From Transformers With Disentangled Attention; DistilBERT: Distilled version of Bidirectional Encoder Representations From Transformers; ELECTRA: Efficiently Learning an Encoder That Classifies Token Replacements Accurately; ERNIE: Enhanced Representation through Knowledge Integration; RoBERTa: Robustly Optimized Bidirectional Encoder Representations From Transformers Pretraining Approach; RUL: right upper lobe.

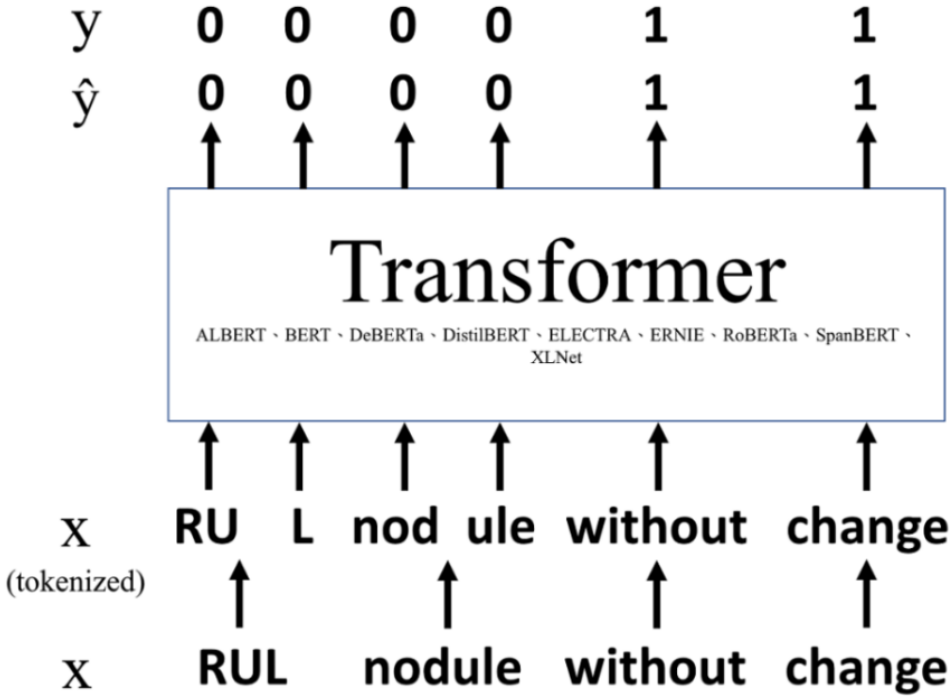
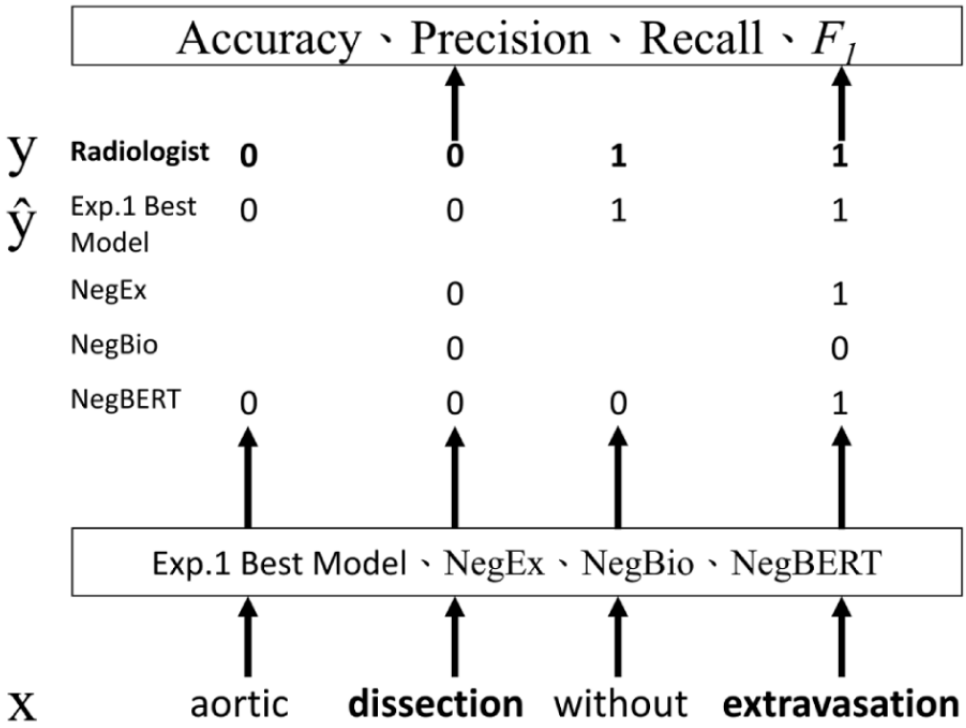


Figure 3. Experiment 2 Note. X: the original text; ŷ: class predicted by the model; y: the gold standard; category 0: positive statements or speculative statements potentially related to abnormal findings; category 1: negative statements or speculative statements unrelated to abnormal findings; bold text: word matching a designated “important keyword.” Exp: experiment.



Modeling in Experiments

The deep learning models used in experiment 1 were ALBERT, BERT, DeBERTa, DistilBERT, ELECTRA, ERNIE, RoBERTa, SpanBERT, and XLNet. All models were fine-tuned based on the pretrained models from Hugging Face.

We used early stopping and used the F_1-score as the model evaluation metric. We used the Adam optimizer with a batch size of 16 and weight decay of 0.01. Table 2 lists the parameters of other models. We set all other unspecified parameters to the default values provided by the open-source PyTorch framework. We segmented the texts into blocks of no more than 510 characters before entering the model to avoid truncation.

We adopted a sequence-to-sequence approach for the training. The training program input the report text in the training and development data set into the model using the corresponding tokenizer and trained the model. The models predicted the token categories using the radiologist-annotated data as the gold standard. The test data set was not included in the training process.

For the NegEx algorithm, we used the negspaCy pipeline component of the open-source Spacy software [41]. The specific named entity recognition model used was "en_ner_bc5cdr_md." In addition, we extended the recognizable entities in Spacy to include all the important keywords defined in our experiment.

We used the previously published training parameters of NegBERT, including a batch size of 8, maximum training epochs of 60, an initial learning rate of 3×10^{-5}, and an early stopping patience of 6. We applied NegBERT for cue detection using the model "bert-base-uncased" and scope detection using the model "xlnet-base-cased." Furthermore, we validated that the trained NegBERT showed a comparable level of performance to that reported in the original publication on the data set specified in the original study.

In addition to the configuration mentioned earlier, we made only minimal modifications to NegBio and NegBERT, such as specifying the dependent software versions, adding the necessary files to the installation, and configuring file paths to ensure the proper execution of the software.

In experiment 2, we optimized the performance of the NegEx, NegBio, and NegBERT methods. This optimization was achieved by modifying the expert-defined rules of NegEx and NegBio and using our training and development data set, as well as the negation and speculation cues we identified, to train NegBERT without using the data set from the original study.

Table 2. Deep learning model and training parameters used in this study.

Model	Learning rate	Warm-up steps	Adam beta1	Adam beta2	Adam epsilon	FP16[a]
ALBERT[b]	1×10^{-5}	10,000	0.9	0.999	1×10^{-8}	False
BERT[c]	1×10^{-4}	10,000	0.9	0.999	1×10^{-8}	False
DeBERTa[d]	1×10^{-4}	10,000	0.9	0.999	1×10^{-6}	True
DistilBERT[e]	2×10^{-5}	0	0.9	0.999	1×10^{-8}	False
ELECTRA[f]	1×10^{-4}	10,000	0.9	0.999	1×10^{-6}	False
ERNIE[g]	5×10^{-5}	4000	0.9	0.98	1×10^{-8}	False
RoBERTa[h]	1×10^{-4}	10,000	0.9	0.999	1×10^{-8}	False
SpanBERT	5×10^{-5}	10,000	0.9	0.999	1×10^{-8}	False
XLNet	2×10^{-5}	10,000	0.9	0.999	1×10^{-6}	False

[a]FP16: half-precision floating-point format.

[b]ALBERT: A Lite Bidirectional Encoder Representations From Transformers.

[c]BERT: Bidirectional Encoder Representations From Transformers.

[d]DeBERTa: Decoding-Enhanced Bidirectional Encoder Representations From Transformers With Disentangled Attention.

[e]DistilBERT: Distilled version of Bidirectional Encoder Representations from Transformers.

[f]ELECTRA: Efficiently Learning an Encoder That Classifies Token Replacements Accurately.

[g]ERNIE: Enhanced Representation through Knowledge Integration.

[h]RoBERTa: Robustly Optimized Bidirectional Encoder Representations From Transformers Pretraining Approach.

Results

Demographics

The data set included in this study consisted of 6000 radiology reports, including plain radiography reports (2538/6000, 42.3%), CT reports (2163/6000, 36.05%), MRI reports (668/6000, 11.13%), ultrasound reports (483/6000, 8.05%), angiography reports (97/6000, 1.62%), and reports from other types of studies (51/6000, 0.85%). The report was completed by 78 radiology residents and their attending physicians. The training, validation and test data sets were mutually exclusive with no overlap in the samples.

The data set used in this study consisted of 78,901 sentences and 704,512 words. A total of 15.01% (105,755/704,512) of all

the words in the data set, were identified as negative and speculative statements unrelated to abnormal findings. Table 3 presents examples and frequencies of these statements. In this study, we defined a "word" as a contiguous sequence of one or more non–white space characters of maximum length. For example, "(−) metastasis" contains 2 words.

Of all the 16,374 cases of sentential negations identified, 15,568 (95.1%) used "no," "without," "not," or "none" as the first word of the negative statement. Furthermore, of all the 2763 cases of negation using symbols or abbreviations, we observed that 2411 (87.2%) used (−), (_), ( ), or [−] at the beginning, end, or middle of the negated clause.

Table 4 presents the frequency and number of occurrences of important keywords, as defined in this study, within negative or speculative statements unrelated to abnormal findings and the total number of occurrences in the study.

Table 3. Types and numbers of negative and the speculative sentences unrelated to abnormal findings included in this study (N=19,467).

Type	Example	Findings, n (%)
Sentential negation	• No evidence of aortic dissection	16,374 (84.11)
Symbols or abbreviations	• Metastasis (−) • Thrombosis: No • DM[a]- HTN[b]- • Anti-HCV[c] [Negative] • - lung - bone	2762 (14.19)
Speculative statements not related to abnormal findings	• No CT[d] evidence of pulmonary embolism. Suggest V/Q[e] scan to exclude small branch embolism if indicated • Metallic artifacts, lesion may be obscured	196 (1.01)
Morphological negation	• This coronary CT scan is nondiagnostic.	135 (0.69)

[a]DM: diabetes mellitus.

[b]HTN: hypertension.

[c]HCV: hepatitis C virus.

[d]CT: computed tomography.

[e]V/Q: ventilation and perfusion.

Table 4. Occurrence and frequency of important keywords defined in this study within negative or the speculative statements unrelated to abnormal findings.

Keywords and their overall occurrences (n=11,480)	Occurrences (N+S)[a], n (%)
Pneumothorax, n=1288 (11.22%)	976 (75.78)
Extravasation, n=182 (1.58%)	84 (46.2)
Fracture, n=2161 (18.82%)	992 (45.90)
Tumor, n=2698 (23.5%)	1025 (37.99)
Infarct, n=1364 (11.88%)	514 (37.68)
Consolidation, n=428 (3.73%)	152 (35.5)
Pneumoperitoneum, n=63 (0.55%)	19 (30)
Thrombosis, n=614 (5.35%)	143 (23.3)
Dissection, n=673 (5.86%)	147 (21.8)
Metastasis, n=1876 (16.34%)	450 (23.98)
Bleeding, n=118 (1.03%)	27 (22.9)
Pulmonary TB[b], n=15 (0.13%)	0 (0)

[a]Number of occurrences within negative or speculative statements unrelated to abnormal findings.

[b]TB: tuberculosis.

Result of Experiment 1

Table 5 presents the results of experiment 1. The accuracy of all transformer-based models included in this experiment was greater than 0.98 for both the training, validation, and test data sets, with macro F_1-scores >0.90. The best-performing model, ALBERT, was selected for inclusion in experiment 2.

Table 5. Comparison of deep learning prediction performance.

	Train and validation data set				Test data set			
	Precision	Recall	F_1	Accuracy	Precision	Recall	F_1	Accuracy
ALBERT[a]	0.992	0.990	0.992	0.998	*0.973* [b]	*0.943* [b]	*0.958* [b]	*0.991* [b]
BERT[c]	0.980	0.986	0.983	0.995	0.960	0.930	0.945	0.989
DeBERTa[d]	0.989	0.971	0.975	0.993	0.958	0.859	0.906	0.980
DistilBERT[e]	0.994	0.990	0.992	0.998	0.980	0.912	0.945	0.988
ELECTRA[f]	0.982	0.982	0.982	0.995	0.956	0.943	0.950	0.989
ERNIE[g]	0.987	0.984	0.986	0.996	0.963	0.920	0.941	0.988
RoBERTa[h]	0.959	0.979	0.969	0.991	0.890	0.933	0.911	0.980
SpanBERT	0.992	0.992	0.992	0.998	0.958	0.932	0.945	0.988
XLNet	0.993	0.993	0.993	0.998	0.970	0.943	0.957	0.990

[a]ALBERT: A Lite Bidirectional Encoder Representations From Transformers.

[b]Italics highlight that the performance of A Lite Bidirectional Encoder Representations From Transformers is the best comparing to the control method across various performance metrics.

[c]BERT: Bidirectional Encoder Representations From Transformers.

[d]DeBERTa: Decoding-Enhanced Bidirectional Encoder Representations From Transformers With Disentangled Attention.

[e]DistilBERT: Distilled version of Bidirectional Encoder Representations from Transformers.

[f]ELECTRA: Efficiently Learning an Encoder That Classifies Token Replacements Accurately.

[g]ERNIE: Enhanced Representation through Knowledge Integration.

[h]RoBERTa: Robustly Optimized Bidirectional Encoder Representations From Transformers Pretraining Approach.

Result of Experiment 2

Before optimization, the performance of NegBio and NegBERT was suboptimal. The F_1-scores for NegEx, NegBio, and NegBERT were 0.889, 0.587, and 0.393, respectively. Our optimization significantly improved the performance of NegBio and NegBERT by increasing their F_1-scores by 0.239 and 0.588, respectively.

Table 6 shows the performance of ALBERT and optimized NegEx, NegBio, and NegBERT. The precision, recall, and F_1-score of our fine-tuned transformer-based model (ALBERT) were better than those of the optimized NegEx, NegBio, and NegBERT.

Table 7 shows the performance evaluation of keyword extraction before and after applying the different negation and speculation-detection algorithms. The ALBERT method resulted in the most significant performance improvement in extracting keywords from positive and speculative statements potentially associated with abnormal findings.

Table 6. Comparison of performance of A Lite Bidirectional Encoder Representations From Transformers (ALBERT) and optimized NegEx, NegBio, and NegBERT in the test data set.

	Precision	Recall	F_1	Accuracy
ALBERT	*0.991*[a]	*0.992*[a]	*0.991*[a]	*0.996*[a]
NegEx	0.886	0.958	0.921	0.959
NegBio	0.860	0.794	0.826	0.917
NegBERT	0.992	0.970	0.981	0.991

[a]Italics highlight that the performance of ALBERT is the best comparing to the control method (NegEx, NegBio, NegBERT) across various performance metrics.

Table 7. Comparison of the performance of keyword extraction in the test data set both before and after applying A Lite Bidirectional Encoder Representations From Transformers (ALBERT) and optimized NegEx, NegBio, and NegBERT.

	Precision	Recall	F_1	Accuracy
ALBERT	*0.998*[a]	*0.997*[a]	*0.997*[a]	*0.996*[a]
NegEx	0.986	0.959	0.972	0.959
NegBio	0.934	0.958	0.945	0.917
NegBERT	0.99	0.998	0.994	0.991
Baseline[b]	0.752	1.00	0.859	0.752

[a]Italics highlight that the performance of ALBERT is the best comparing to the control method (NegEx, NegBio, NegBERT) and baseline (no negation or speculation detection were performed) across various performance metrics.

[b]All named entities considered "positive." No negation or speculation-detection algorithm was applied.

Sources of Errors

Overview

We analyzed the sources of the errors (Table 8). Despite changes in the rules defined by the experts, errors persisted in NegEx and NegBio. We identified the following causes:

Table 8. Analysis of the causes of errors in different methods (after optimization).

Method and cause of the wrong prediction[a]	Counts, n (%)
NegBio (n=177)	
Errors in the extraction of named entities	58 (32.8)
Symbol-related errors	49 (27.7)
Tokenization error	21 (11.9)
Errors in the prediction of speculative statements	14 (7.9)
NegEx (n=87)	
False-positive prediction related to speculative statements	37 (42)
Trigger word not triggered	21 (24)
Incorrect scope resolution	16 (18)
Symbol-related errors	6 (6)
NegBERT (n=20)	
All false-negative predictions	16 (80)
All false-positive predictions	4 (20)
False-positive predictions related to speculative statements unrelated to abnormal findings	0 (0)
ALBERT[b] (n=9)	
All false-positive predictions	5 (55)
False-positive predictions related to speculative statements unrelated to abnormal findings	0 (0)

[a]The table only list the most important causes of identifiable error.

[b]ALBERT: A Lite Bidirectional Encoder Representations From Transformers.

Findings of NegEx

First, we found many errors owing to incompatibility between the NegEx method for identifying speculative statements and the study requirements. NegEx made identical predictions for all keywords in the identified speculative statements regardless of their relevance to abnormal findings. However, our study categorized keywords in speculative sentences differently based on their relevance to abnormal findings, leading to discrepancies with NegEx's results.

Second, the trigger word would only sometimes trigger. For example, in the phrase "1.No evidence of tumor," the trigger word "No" would not be recognized because it was concatenated with the character "1." without any intervening space.

Third, errors also occurred owing to the misinterpretation of the scope of negation and speculation, such as misinterpreting "No improvement of the tumor" as "No tumor."

Fourth, errors occurred in the presence of symbols in radiology reports; for example, the use of special symbols by radiologists that are undefined in the trigger word or the confusion caused by the co-occurrence of special symbols that express a positive and a negative statement: (−) fatty liver and (+) portal vein thrombosis.

Findings of NegBio

We identified the following errors when using NegBio:

First, errors occurred in named entity extraction. The named entities in NegBio's output file might be missing target keywords or had incorrect positions, resulting in incorrect future analyses.

Second, errors occurred when the radiology report contained negations using symbols or abbreviations, such as "metastasis (−)." Our analysis showed that these symbols could lead to unpredictable results in syntactic structure analysis and subsequent analyses.

Third, combining words with numerals or punctuation marks leads to errors in tokenization and subsequent analysis. For example, "1.No" in "1.No obvious acute infarct or brain metastasis" was not correctly parsed as "No."

Fourth, many errors occurred because NegBio made identical predictions for diagnostic keywords in all speculative sentences, regardless of their relevance to abnormal findings. This behavior was inconsistent with the labeling of this experiment.

Findings of NegBERT and ALBERT

We observed the suboptimal performance of NegBERT when applied to corpora from different domains and tasks. The performance of NegBERT trained on the Simon Fraser University review corpus was suboptimal when evaluated on our corpus and task. Retraining NegBERT with our data significantly improved its performance, indicating that the poor performance was primarily due to differences in the training data and labeling.

Our error analysis showed that retrained NegBERT and ALBERT made fewer errors than the other methods in predicting whether words occurred in speculative statements unrelated to abnormal findings. The number of all false-positive predictions by NegBERT and ALBERT was 4 and 5, respectively. Both were lower than the number of false-positive predictions made by NegEx and NegBio for this prediction task, indicating higher specificity. However, because we grouped all negative and speculative statements not related to abnormal findings into the same category, we could not calculate the exact value of specificity. Both models showed 100% sensitivity in identifying important diagnostic keywords in speculative statements unrelated to abnormal findings, with no false-negative predictions.

Owing to the complexity of BERT, we could not further analyze the causes of other errors.

Discussion

Principal Findings

Overview

This study found that 39.45% (4529/11,480) of the important diagnostic keywords occurred in negative or speculative statements unrelated to abnormal findings, posing a challenge for automatic labeling by LISs and information extraction techniques.

Our study proposes a deep learning method that accurately distinguishes whether diagnostic keywords are in negative or speculative statements unrelated to abnormal findings. Our research has revealed the shortcomings of existing methods, including NegEx, NegBio, and NegBERT, while highlighting the advantages of our proposed approach over these methods.

Limitation of NegEx and NegBio

We observed common errors in Spacy's NegEx and NegBio that the expert rule adjustment could not resolve.

First, several vital errors in NegEx and NegBio, including errors related to trigger words in NegEx, tokenization errors in NegBio, and symbol-related errors in NegEx and NegBio, were attributed to interference from punctuation and numerals. For example, in the radiology reports in our sample, English sentences were often combined with numbers and punctuation marks and written as numbered or bulleted lists, such as "1.No evidence of aortic dissection" In addition, using symbols or abbreviations in the form of checklists was also common. For example, "Metastasis (−)" or "Anti-HCV [Negative]" were frequently used. Our results showed that NegEx and NegBio could not handle this issue correctly.

Second, NegEx and NegBio also caused many errors in the analyses where the simultaneous observation of multiple sentences is required. Our data showed that it is often necessary to examine multiple sentences simultaneously to determine whether speculative statements are associated with abnormal findings. For example, in "No CT evidence of large infarct. Suggest MRI to exclude hyperacute infarct if indicated," without considering the first sentence, which denies the finding of infarct evidence, it cannot be determined that the "hyperacute infarct" in the second sentence is unrelated to the actual findings. NegEx and NegBio, which are designed to analyze sentences in isolation without considering contextual information, cannot meet this requirement.

Our results regarding NegEx are consistent with previous research of Wu et al [42], highlighting the importance of tuning algorithms such as NegEx to achieve optimal performance in different corpora. Our results also confirm that NegEx produces incorrect results owing to improper negation scope resolution [22].

We found that NegBio requires modifying expert-defined rules to improve its performance. Our study is the first to report NegBio's limited generalizability in real-world radiology reports across all body parts. We also observed problems with the implementation of NegBio.

Limitation of NegBERT

Our experiment showed a significant improvement in NegBERT's performance after retraining on our hospital data set. The difference in the training data and annotations is likely the reason for the initial poor performance of NegBERT.

This observation is consistent with previous findings that deep learning models such as BERT tend to perform poorly on out-of-domain corpora. For example, a study by Miller et al [39] using RoBERTa for negation detection on both in-domain and out-of-domain corpora observed F_1-scores of 0.95 and 0.583, respectively. Our experiment supports this result and shows that the drop in F_1-scores can be even worse depending on the corpus and task.

Advantages of ALBERT and BERT Transformer

We performed a comparison between the ALBERT and NegBERT methods and made the following key observations.

First, learning the negation cue and scope in 2 steps provides a limited performance improvement. Our method takes a different approach from NegBERT and traditional negation recognition studies in that our model learn the entire part of the sentence containing both the cue and scope in the same step without explicitly telling the model which word is the "cue" of the negation or speculation. However, the performance was still better than that of the retrained NegBERT. The study by Sergeeva et al [11] based on LSTM suggests that the deep learning method can learn negation cue information to some extent automatically, with performance comparable with that of automatic cue prediction algorithms. Our results show that BERT might have a similar capability. Our results suggest that providing additional cue information through expert annotation may not significantly improve performance compared with other factors, such as model selection, hyperparameter optimization, and training techniques.

Second, our results show that the model size and complexity do not necessarily correlate with improved performance. In our study, the fine-tuned ALBERT model outperformed larger and more complex models, including BERT and XLNet used by NegBERT, as well as RoBERTa used in the study by Miller et al [39]. The use of lightweight models, such as ALBERT, may have practical advantages, including reduced computational resource requirements and training time, compared with BERT [31].

In our study, ALBERT and retrained NegBERT outperformed NegEx and NegBio in terms of the number of false-positive predictions and specificity while maintaining 100% sensitivity in predicting whether keywords occurred in speculative sentences unrelated to abnormal findings. This task required multisentence context analysis of our data set, and our results suggest that BERT can look at multiple sentences simultaneously. The attention mechanism is a reasonable explanation for this phenomenon.

Comparison With Prior Work

Our study fine-tuned the ALBERT model using a more comprehensive data set that included a broader range of imaging modalities and subspecialties than previous studies. Table 9 shows the best performances and corresponding data sets used in previous studies that detected whether named entities occurred in negation and speculation in radiology reports. The range of imaging modalities and subspecialties represented in the radiology reports in these studies was limited, such as chest x-ray reports only in the study by Peng et al [23] or brain CT and MRI reports only in the studies by Grivas et al [43] and Sykes et al [12]. We hypothesized that including a more diverse set of examination and imaging subspecialties in the data results in a more representative sample of the report content and improves the model's generalizability. Our results support this hypothesis, as the ALBERT model showed only a 0.034 decrease in its F_1-score on an unseen test data set with different disease types and inputs from different physicians.

Our experiments also address a more difficult speculation-detection task than previous studies; however, ALBERT still demonstrates good performance. This distinction requires the ability of the algorithm to consider multiple sentences simultaneously in our data set. To the best of our knowledge, our study is the first to propose a distinction between speculative sentences related and unrelated to abnormal findings based on the application scenario to facilitate more precise filtering and the first study to highlight the impact of the lack of multisentence analysis in negation detection algorithms.

Table 9. Comparison of best performances between studies distinguishing whether named entities occurred in negation or speculation.

Study	Algorithm	P[a]	R[b]	F_1	Best[c]	N[d]	Task[e]	Type[f]
Our study	ALBERT[g]	0.991	0.992	0.991	Test data set	6000	ND[h]+S*[i]	All body parts
Sykes et al [12]	BiLSTM[j]	0.973	0.981	0.977	ESS[k]	630	ND+S[l]	Brain CT[m] and MRI[n]
Peng et al [23]	NegBio	0.944	0.944	0.944	Chest x-ray	900	ND+S	Chest x-ray
Grivas et al [43]	Edie-R	0.925	0.943	0.934	ESS	630	ND+S	Brain CT and MRI

[a]P: precision.

[b]R: recall.

[c]Name of the best-performing data set. Other data sets are not included.

[d]Number of samples in the best-performing data set; other data sets are not included.

[e]Task performed in the study.

[f]Types of radiologic studies included in the study.

[g]ALBERT: A Lite Bidirectional Encoder Representations From Transformers.

[h]ND: negation detection.

[i]S*: detection of speculation unrelated to abnormal findings.

[j]BiLSTM: Bidirectional Long Short-Term Memory.

[k]ESS: Edinburgh Stroke Study.

[l]S: speculation detection.

[m]CT: computed tomography.

[n]MRI: magnetic resonance imaging.

Implication in Clinical Practice

We found problems with NegEx and NegBio in that modifying expert-defined rules could not be solved, including difficulties with numbers and punctuation, implementation-specific challenges, and the design constraint of observing only a single sentence at a time; thus, NegEx and NegBio should be used cautiously or avoided in such situations to prevent errors. On the basis of our data, we also found that NegBio and NegBERT have limitations in generalizability, making them inappropriate for use without training or modeling.

Our results indicate that BERT is more suitable than NegEx and NegBio for tasks involving multisentence context analysis, similar to the experiment conducted in this study. NegEx and NegBio were designed for single-sentence analysis because they segmented the text into independent sentences. This approach limits the ability to incorporate contextual information from other sentences into the analysis. While NegEx and NegBio can perform binary classification of words in sentences as speculative or not, they lack the capacity for further granular differentiation based on contextual information.

We found that the training process of the transformers did not require 2 separate learning phases for cue and scope. Our findings could reduce the workload of expert annotation in clinical applications, as the explicit annotation of cues in a separate step requires additional work. This hypothesis needs further testing in future studies.

Our results show that deep learning models outperform non–deep learning methods, and lightweight models such as ALBERT can achieve superior performance and outperform other transformer-based models. However, fine-tuning based on the specific domain corpus and task is still essential regardless of the model used.

Limitations

The data were obtained from 3 internal branches of a single institution and not from publicly available data sets. In addition, the speculation-detection task differed from previous studies in this area. The comparability of the performance with that of previous studies may be limited. If open data using the same annotation methodology become available, subsequent research could verify our findings by implementing the same model on the open data set.

Our study optimized the control methods (NegEx and NegBio), but we cannot exclude the possibility of further performance improvement by modifying or adding expert rules. However, this highlights the limitations of an expert rule–based approach, which requires experts not only to detect negations and speculations but also to summarize and modify rules manually. Moreover, expert rules cannot resolve the algorithmic design or implementation constraints.

To prevent the deep learning model from training failure, we combined negative statements with speculative statements unrelated to abnormal findings in the same category because of the low proportion of the latter. As a result, we cannot separately evaluate the model's performance on negative and speculative sentences unrelated to abnormal findings or accurately quantify the latter's performance. Nevertheless, metrics such as the number of false-positive predictions can still be used to compare the performance between methods.

Conclusions

Manual free-text reporting remains the norm in radiology worldwide, hampering the ability to perform computer-assisted

analyses. The presence of information irrelevant to the actual findings poses a significant challenge to the implementation of automatic radiology report highlighting, flagging, or information extraction.

Previous research on negation and speculation detection in radiology has aimed to identify all instances. Our study advances this by targeting only speculative statements unrelated to abnormal findings and improving the discrimination of relevant information using BERT's multisentence contextual analysis capabilities.

Lightweight transformer models, such as ALBERT, can outperform NegEx, NegBio, and NegBERT on more complex and diverse real-world radiology reports. Despite achieving good results on public data sets, NegBio and NegBERT demonstrated different performances on more complicated real-world radiology reports.

Our research has potential applications in academia and clinical practice. Future studies may consider including lightweight models such as ALBERT. In clinical practice, our method achieved high performance. It can help algorithms such as keyword highlighting in hospital information systems to identify passages of potentially important information without false alarms, improving physician efficiency and health care quality. Our results also apply to radiology report information retrieval, such as search engines, in which negative and speculative statements unrelated to abnormalities can lead to incorrect results.

Acknowledgments

This research received no specific grants from any funding agency in the public, commercial, or not-for-profit sectors.

Authors' Contributions

KHW proposed the research topic and experimental design and completed the recruitment, data analysis, computer programming, and writing of the entire paper.
CFL contributed to research design improvement and manuscript proofreading.
CJC performed big medical data exporting and cleaning and manuscript proofreading.

Conflicts of Interest

None declared.

References

1. Lacson R, Prevedello LM, Andriole KP, O'Connor SD, Roy C, Gandhi T, et al. Four-year impact of an alert notification system on closed-loop communication of critical test results. AJR Am J Roentgenol 2014 Dec;203(5):933-938 [FREE Full text] [doi: 10.2214/AJR.14.13064] [Medline: 25341129]
2. Ignácio FC, de Souza LR, D'Ippolito G, Garcia MM. Radiology report: what is the opinion of the referring physician? Radiol Bras 2018 Sep;51(5):308-312 [FREE Full text] [doi: 10.1590/0100-3984.2017.0115] [Medline: 30369658]
3. Reda AS, Hashem DA, Khashoggi K, Abukhodair F. Clinicians' behavior toward radiology reports: a cross-sectional study. Cureus 2020 Nov 05;12(11):e11336 [FREE Full text] [doi: 10.7759/cureus.11336] [Medline: 33304672]
4. European Society of Radiology (ESR). ESR guidelines for the communication of urgent and unexpected findings. Insights Imaging 2012 Feb;3(1):1-3 [FREE Full text] [doi: 10.1007/s13244-011-0135-y] [Medline: 22695992]
5. Nakamura Y, Hanaoka S, Nomura Y, Nakao T, Miki S, Watadani T, et al. Automatic detection of actionable radiology reports using bidirectional encoder representations from transformers. BMC Med Inform Decis Mak 2021 Sep 11;21(1):262 [FREE Full text] [doi: 10.1186/s12911-021-01623-6] [Medline: 34511100]
6. Perrotta PL, Karcher DS. Validating laboratory results in electronic health records: a College of American Pathologists Q-Probes study. Arch Pathol Lab Med 2016 Sep;140(9):926-931 [FREE Full text] [doi: 10.5858/arpa.2015-0320-CP] [Medline: 27575266]
7. Srinivasa Babu A, Brooks ML. The malpractice liability of radiology reports: minimizing the risk. Radiographics 2015 Mar;35(2):547-554. [doi: 10.1148/rg.352140046] [Medline: 25763738]
8. Ruskin KJ, Hueske-Kraus D. Alarm fatigue: impacts on patient safety. Curr Opin Anaesthesiol 2015 Dec;28(6):685-690. [doi: 10.1097/ACO.0000000000000260] [Medline: 26539788]
9. Lazib L, Zhao Y, Qin B, Liu T. Negation scope detection with recurrent neural networks models in review texts. In: Proceedings of the 2nd International Conference of Young Computer Scientists, Engineers and Educators: Social Computing. 2016 Presented at: ICYCSEE' 16; August 20-22, 2016; Harbin, China p. 494-508.
10. Taylor SJ, Harabagiu SM. The role of a deep-learning method for negation detection in patient cohort identification from electroencephalography reports. AMIA Annu Symp Proc 2018 Dec 05;2018:1018-1027 [FREE Full text] [Medline: 30815145]
11. Sergeeva E, Zhu H, Prinsen P, Tahmasebi A. Negation scope detection in clinical notes and scientific abstracts: a feature-enriched LSTM-based approach. AMIA Jt Summits Transl Sci Proc 2019 May 06;2019:212-221 [FREE Full text] [Medline: 31258973]

12. Sykes D, Grivas A, Grover C, Tobin R, Sudlow C, Whiteley W, et al. Comparison of rule-based and neural network models for negation detection in radiology reports. Nat Lang Eng 2021 Mar;27(2):203-224 [FREE Full text] [doi: 10.1017/s1351324920000509]
13. Khandelwal A, Britto BK. Multitask learning of negation and speculation using transformers. In: Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis. 2020 Presented at: Louhi' 20; November 20, 2020; Virtual p. 79-87. [doi: 10.18653/v1/2020.louhi-1.9]
14. Fancellu F, Lopez A, Webber B. Neural networks for negation scope detection. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. 2016 Presented at: ACL' 16; August 7-12, 2016; Berlin, Germany p. 495-504. [doi: 10.18653/v1/P16-1047]
15. Gautam D, Maharjan N, Banjade R, Tamang LJ, Rus V. Long short term memory based models for negation handling in tutorial dialogues. In: Proceedings of the 31st International Florida Artificial Intelligence Research Society Conference. 2018 Presented at: FLAIRS' 18; May 21-23, 2018; Melbourne, FL, USA p. 14-19. [doi: 10.13140/RG.2.2.26250.36804]
16. Khandelwal A, Sawant S. NegBERT: a transfer learning approach for negation detection and scope resolution. In: Proceedings of the 12th Language Resources and Evaluation Conference. 2020 Presented at: LREC' 20; May 11-16, 2020; Marseille, France p. 5739-5748 URL: https://aclanthology.org/2020.lrec-1.704.pdf
17. Qian Z, Li P, Zhu Q, Zhou G, Luo Z, Luo W. Speculation and negation scope detection via convolutional neural networks. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. 2016 Nov Presented at: EMNLP' 16; November 1-5, 2016; Austin, TX, USA p. 815-825. [doi: 10.18653/v1/D16-1078]
18. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform 2001 Oct;34(5):301-310 [FREE Full text] [doi: 10.1006/jbin.2001.1029] [Medline: 12123149]
19. Mutalik PG, Deshpande A, Nadkarni PM. Use of general-purpose negation detection to augment concept indexing of medical documents: a quantitative study using the UMLS. J Am Med Inform Assoc 2001 Nov;8(6):598-609 [FREE Full text] [doi: 10.1136/jamia.2001.0080598] [Medline: 11687566]
20. Gindl S, Kaiser K, Miksch S. Syntactical negation detection in clinical practice guidelines. Stud Health Technol Inform 2008;136:187-192 [FREE Full text] [Medline: 18487729]
21. Aronow DB, Fangfang F, Croft WB. Ad hoc classification of radiology reports. J Am Med Inform Assoc 1999 Sep;6(5):393-411 [FREE Full text] [doi: 10.1136/jamia.1999.0060393] [Medline: 10495099]
22. Mehrabi S, Krishnan A, Sohn S, Roch AM, Schmidt H, Kesterson J, et al. DEEPEN: a negation detection system for clinical text incorporating dependency relation into NegEx. J Biomed Inform 2015 Apr;54:213-219 [FREE Full text] [doi: 10.1016/j.jbi.2015.02.010] [Medline: 25791500]
23. Peng Y, Wang X, Lu L, Bagheri M, Summers R, Lu Z. NegBio: a high-performance tool for negation and uncertainty detection in radiology reports. AMIA Jt Summits Transl Sci Proc 2018 May 18;2017:188-196 [FREE Full text] [Medline: 29888070]
24. Medlock B, Briscoe T. Weakly supervised learning for hedge classification in scientific literature. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. 2007 Presented at: ACL' 07; June 25-27, 2007; Prague, Czech Republic p. 992-999 URL: https://aclanthology.org/P07-1125.pdf
25. Rokach L, Romano R, Maimon O. Negation recognition in medical narrative reports. Inf Retr 2008 Jun 7;11(6):499-538 [FREE Full text] [doi: 10.1007/s10791-008-9061-0]
26. Morante R, Liekens A, Daelemans W. Learning the scope of negation in biomedical texts. In: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing. 2008 Presented at: EMNLP' 08; October 25-27, 2008; Honolulu HI, USA p. 715-724 URL: https://aclanthology.org/D08-1075.pdf [doi: 10.3115/1613715.1613805]
27. Morante R, Daelemans W. A metalearning approach to processing the scope of negation. In: Proceedings of the 13th Conference on Computational Natural Language Learning. 2009 Presented at: CoNLL '09; June 4-5, 2009; Boulder, CO, USA p. 21-29. [doi: 10.3115/1596374.1596381]
28. Ou Y, Patrick J. Automatic negation detection in narrative pathology reports. Artif Intell Med 2015 May;64(1):41-50. [doi: 10.1016/j.artmed.2015.03.001] [Medline: 25990897]
29. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019 Presented at: NAACL' 19; June 2-7, 2019; Minneapolis, MN, USA p. 4171-4186. [doi: 10.18653/v1/N19-1423]
30. Rivera Zavala R, Martinez P. The impact of pretrained language models on negation and speculation detection in cross-lingual medical text: comparative study. JMIR Med Inform 2020 Dec 03;8(12):e18953 [FREE Full text] [doi: 10.2196/18953] [Medline: 33270027]
31. Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R. ALBERT: a lite BERT for self-supervised learning of language representations. In: Proceedings of the 2020 International Conference on Learning Representations. 2020 Presented at: ICLR '20; April 26-30, 2020; Addis Ababa, Ethiopia URL: https://openreview.net/pdf?id=H1eA7AEtvS

32. Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In: Proceedings of the 2019 Conference on Neural Information Processing Systems. 2019 Presented at: NeurIPS '19; December 8-14, 2019; Vancouver, Canada.
33. Yang Z, Dai Z, Yang Y, Carbonell JG, Salakhutdinov RR, Le QV. Xlnet: generalized autoregressive pretraining for language understanding. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. 2019 Presented at: NIPS'19; December 8-14, 2019; Vancouver, Canada p. 5753-5763 URL: https://dl.acm.org/doi/10.5555/3454287.3454804
34. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: a robustly optimized BERT pretraining approach. In: Proceedings of the 2020 International Conference on Learning Representations. 2020 Presented at: ICLR '20; April 26-30, 2020; Addis Ababa, Ethiopia URL: https://openreview.net/attachment?id=SyxS0T4tvS&name=original_pdf
35. Sun Y, Wang S, Li Y, Feng S, Chen X, Zhang H, et al. ERNIE: enhanced representation through knowledge integration. arXiv Preprint posted online on April 19, 2018. [doi: 10.48550/arXiv.1904.09223]
36. Joshi M, Chen D, Liu Y, Weld DS, Zettlemoyer L, Levy O. SpanBERT: improving pre-training by representing and predicting spans. Trans Assoc Comput Linguist 2020 Dec;8:64-77 [FREE Full text] [doi: 10.1162/tacl_a_00300]
37. He P, Liu X, Gao J, Chen W. DeBERTa: decoding-enhanced BERT with disentangled attention. In: Proceedings of the 2021 International Conference on Learning Representations. 2021 Presented at: ICLR '21; May 3-7, 2021; Virtual URL: https://openreview.net/pdf?id=XPZIaotutsD
38. Clark K, Luong MT, Le QV, Manning CD. Electra: pre-training text encoders as discriminators rather than generators. In: Proceedings of the 8th International Conference on Learning Representations. 2020 Presented at: ICLR' 20; April 26-May 1, 2020; Virtual p. 1-18 URL: https://openreview.net/pdf?id=r1xMH1BtvB
39. Miller T, Laparra E, Bethard S. Domain adaptation in practice: lessons from a real-world information extraction pipeline. In: Proceedings of the 2nd Workshop on Domain Adaptation for NLP. 2021 Apr 20 Presented at: AdaptNLP' 21; April 20, 2021; Kyiv, Ukraine p. 105-110 URL: https://aclanthology.org/2021.adaptnlp-1.11.pdf
40. Hiroki N, Takahiro K, Junya K, Yasufumi T, Xu L. doccano: text annotation tool for human. GitHub. 2018. URL: https://github.com/doccano/doccano [accessed 2022-11-15]
41. Honnibal M, Montani I, Van Landeghem S, Boyd A. spaCy: industrial-strength natural language processing in Python. spaCy. 2020. URL: https://spacy.io [accessed 2022-11-15]
42. Wu S, Miller T, Masanz J, Coarr M, Halgrim S, Carrell D, et al. Negation's not solved: generalizability versus optimizability in clinical natural language processing. PLoS One 2014 Nov 13;9(11):e112774 [FREE Full text] [doi: 10.1371/journal.pone.0112774] [Medline: 25393544]
43. Grivas A, Alex B, Grover C, Tobin R, Whiteley W. Not a cute stroke: analysis of rule- and neural network-based information extraction systems for brain radiology reports. In: Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis. 2020 Presented at: Louhi' 20; November 20, 2020; Virtual p. 24-37 URL: https://aclanthology.org/2020.louhi-1.4.pdf [doi: 10.18653/v1/2020.louhi-1.4]

Abbreviations

ALBERT: A Lite Bidirectional Encoder Representations From Transformers
BERT: Bidirectional Encoder Representations From Transformers
BiLSTM: Bidirectional Long Short-Term Memory
CRF: Conditional Random Field
CT: computed tomography
DeBERTa: Decoding-Enhanced Bidirectional Encoder Representations From Transformers With Disentangled Attention
DEEPEN: Dependency Parser Negation
DistilBERT: Distilled version of Bidirectional Encoder Representations From Transformers
ELECTRA: Efficiently Learning an Encoder That Classifies Token Replacements Accurately
ERNIE: Enhanced Representation through Knowledge Integration
LIS: laboratory information system
LSTM: Long Short-Term Memory
MRI: magnetic resonance imaging
RoBERTa: Robustly Optimized Bidirectional Encoder Representations From Transformers Pretraining Approach

Edited by C Lovis; submitted 08.02.23; peer-reviewed by CJ Lin, D Hu, PP Zhao, S Kim; comments to author 11.03.23; revised version received 21.03.23; accepted 24.03.23; published 25.04.23.

Please cite as:
Weng KH, Liu CF, Chen CJ
Deep Learning Approach for Negation and Speculation Detection for Automated Important Finding Flagging and Extraction in Radiology Report: Internal Validation and Technique Comparison Study
JMIR Med Inform 2023;11:e46348
URL: https://medinform.jmir.org/2023/1/e46348
doi:10.2196/46348
PMID:37097731

Original Paper

Understanding Views Around the Creation of a Consented, Donated Databank of Clinical Free Text to Develop and Train Natural Language Processing Models for Research: Focus Group Interviews With Stakeholders

Natalie K Fitzpatrick[1], MSc; Richard Dobson[2], PhD; Angus Roberts[2], PhD; Kerina Jones[3], PhD; Anoop D Shah[1,4], PhD; Goran Nenadic[5], PhD; Elizabeth Ford[6], PhD

[1]Institute of Health Informatics, University College London, London, United Kingdom
[2]Department of Biostatistics and Health Informatics, King's College London, London, United Kingdom
[3]Department of Population Data Science, Swansea University Medical School, Swansea, United Kingdom
[4]University College London Hospitals NHS Foundation Trust, London, United Kingdom
[5]Department of Computer Science, University of Manchester, Manchester, United Kingdom
[6]Brighton and Sussex Medical School, Brighton, United Kingdom

Corresponding Author:
Natalie K Fitzpatrick, MSc
Institute of Health Informatics
University College London
222 Euston Road
London, NW1 2DA
United Kingdom
Phone: 44 7808032697
Email: n.fitzpatrick@ucl.ac.uk

Abstract

Background: Information stored within electronic health records is often recorded as unstructured text. Special computerized natural language processing (NLP) tools are needed to process this text; however, complex governance arrangements make such data in the National Health Service hard to access, and therefore, it is difficult to use for research in improving NLP methods. The creation of a donated databank of clinical free text could provide an important opportunity for researchers to develop NLP methods and tools and may circumvent delays in accessing the data needed to train the models. However, to date, there has been little or no engagement with stakeholders on the acceptability and design considerations of establishing a free-text databank for this purpose.

Objective: This study aimed to ascertain stakeholder views around the creation of a consented, donated databank of clinical free text to help create, train, and evaluate NLP for clinical research and to inform the potential next steps for adopting a partner-led approach to establish a national, funded databank of free text for use by the research community.

Methods: Web-based in-depth focus group interviews were conducted with 4 stakeholder groups (patients and members of the public, clinicians, information governance leads and research ethics members, and NLP researchers).

Results: All stakeholder groups were strongly in favor of the databank and saw great value in creating an environment where NLP tools can be tested and trained to improve their accuracy. Participants highlighted a range of complex issues for consideration as the databank is developed, including communicating the intended purpose, the approach to access and safeguarding the data, who should have access, and how to fund the databank. Participants recommended that a small-scale, gradual approach be adopted to start to gather donations and encouraged further engagement with stakeholders to develop a road map and set of standards for the databank.

Conclusions: These findings provide a clear mandate to begin developing the databank and a framework for stakeholder expectations, which we would aim to meet with the databank delivery.

(JMIR Med Inform 2023;11:e45534) doi:10.2196/45534

KEYWORDS

consent; databank; electronic health records; free text; governance; natural language processing; public involvement; unstructured text

Introduction

Background

Electronic health records (EHRs) contain a rich narrative of the patient journey and have huge potential for research [1]. However, research using EHRs is typically limited to the structured data (such as numerical values and diagnoses coded using a controlled vocabulary), despite a large proportion of the information in EHRs being in the form of unstructured (free) text. The analysis of free text at scale requires specialized tools and methods (natural language processing [NLP]) to "read," process, and structure the information before it can be used at scale for research purposes.

NLP of clinical text has many potential benefits, both for individual care and improving health services [2]. These include (1) to facilitate the process of clinical coding [3], which is the process by which clinical coding staff in hospitals assign codes from a specific terminology (eg, International Classification of Diseases, 10th revision [ICD-10]) [4] to patient episodes for reimbursement; (2) to facilitate structured recording of diagnoses in clinical care using Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) [5], which is currently not done consistently [6]; and (3) to enable research using information in EHRs, which are currently not coded. Compared with manual review of free text, automated analysis is much faster and enables a much larger amount of text to be analyzed, enabling larger and more representative patient samples to be used for research.

The Challenge

Tools and platforms for text access and analysis such as CogStack (developed by a consortia of scientists at King's College London, King's College Hospital, South London and the Maudsley, Guy's and St Thomas's hospital, University College London, and University College London Hospitals National Health Service [NHS] Foundation Trust and some members of the CogStack open-source community) [7] have been developed and installed at some NHS sites with great success, but overall, access to free text for researchers is still currently difficult. Ideally, free text needs to be brought out of the NHS environment so expert computer scientists working in university research or other non-NHS environments can use the data to train their computer algorithms to extract the important clinical information. In the United Kingdom, the application of NLP for health care text research is largely limited to within large NHS hospital trusts with academic affiliations and in-house NLP expertise owing to complex governance requirements arising from increased concerns around the potential risk of reidentifying patients. In Scotland, a successful model adopted by groups including the Health Informatics Centre at the University of Dundee [8], DataLoch [9], and the national electronic Data Research and Innovation Service [10] in collaboration with the University of Edinburgh Clinical NLP Research Group [11] involves the provision of data for research through secure trusted research environments outside NHS or university settings. Nind et al [12] describe an approach for extracting, linking, deidentifying, and hosting clinical imaging data within a controlled secure environment as a resource for national and international research. This model provides a potential alternative approach to hosting the databank outside of an NHS or university setting and allows for timely and secure access to data; however, the governance framework is complex. In the United Kingdom, before medical data can be shared outside of the NHS environment, identifiers such as names of patients, family members, and health care professionals; addresses; and dates of birth, which can occur anywhere in the text, first need to be removed—a process known as "deidentification" [13]. Even when deidentified, there remains a risk that some identifiable information may have been missed, third parties might be identified, or the narrative may be too revealing.

Routinely collected health data are legally accessed for secondary purposes such as research by 2 lawful bases under UK data protection law: one is the principle of informed consent from the patient and the other is "task in the public interest" [14]. For processing under the lawful basis of "task in the public interest," health care data needs to be deidentified or anonymized before it can be shared outside the clinical environment under the General Data Protection Regulation principle of data minimization [15]. This is where governance becomes difficult, as deidentifying free-text clinic notes, letters, and reports is complex and a rapidly evolving field, and the accuracy of the process is hard to assess [2]. Technology exists to automatically redact identifying information so that only deidentified documents are presented to computer scientists to develop NLP [1]. However, the reidentification risk from automatically deidentified text remains unknown. Many independent health research ethics committees (RECs) do not have the specialized technical knowledge needed to evaluate the risks posed to patients by this type of research, and indeed, many researchers and data custodians are not sure of the scrutiny and approvals needed to legally process free text for research. As a consequence, a conservative approach is usually taken, resulting in heavy restrictions on data access [16]. Therefore, there are currently very few health care free-text data sets available to NLP researchers to develop and evaluate their algorithms; 1 example is the Medical Information Mart for Intensive Care (MIMIC) database [17,18] in the United States. MIMIC is based on a selected patient population (critical care patients) from 1 US center and contains structured and unstructured (eg, diagnostic reports and physician notes) deidentified data linked to hospital EHR and mortality data. Data that will contribute to the databank remain to be decided and will follow further consultation with stakeholders, but at the very minimum, it will include unstructured text from primary care and hospital records for a defined population of patients in the United Kingdom.

The Solution

One possibility for breaking down barriers to access to free-text data is to enable access to clinical text via the lawful basis of informed consent. Creating a "donated" databank of clinical free text in which each patient represented has given informed and explicit consent for their data to be used in this way could provide an important and timely opportunity for NLP researchers to develop and train NLP algorithms to process the free text, which can then be used on other data sets in the NHS to conduct clinical research. NLP researchers in universities or other non-NHS settings only need to access a sample of patient free-text data to develop and train the NLP algorithms on the databank, which could then be run on unseen patient free text housed within the NHS for research, avoiding the important privacy issues laid out above.

To test early thinking on the databank, we carried out a series of in-depth focus groups among 4 key stakeholder groups to find out what key stakeholders think about a consented, donated databank of clinical free text to help create, train, and evaluate NLP for clinical research.

Methods

Participant Selection and Inclusion Criteria

Four stakeholder groups were identified based on their potential interest and investment in the databank as follows: (1) patients and members of the public, (2) clinicians (NHS general practitioners [GPs], hospital doctors, and doctors in training), (3) NHS Trust information governance (IG) leads and REC members, and (4) NLP researchers based in universities or NHS hospitals. Participants lived in the United Kingdom and were aged ≥18 years. Patients and members of the public were based in the community and had to have some previous knowledge or understanding of the use of free-text health data for research; for example, they may have attended events or workshops on this topic or had experience participating in advisory committees on the use of free text.

Participant Recruitment

Patients and members of the public were recruited via an advert posted by existing networks, including Health Data Research United Kingdom [19] and the National Institute for Health Research People in Research network [20]. Other stakeholders were identified via existing national networks, contacts, and organizations and were approached directly by email by the research team. In addition, IG leads were identified via the Office of the National Data Guardian [21] and by searching the websites of NHS Trusts and Health Boards. NHS Research Ethics Service committee members were identified by searching the NHS Health Research Authority website [22], and academic NLP researchers were identified via professional networks including the UK health care text analytics network known as Healtex [23].

Potential participants were invited by the research team by email to participate in 1 of the 4 relevant stakeholder focus groups. Before deciding whether to take part, participants were asked to read a study information sheet and return a completed expression of interest form recording basic demographic information including age category (deciles), gender identity, and ethnicity. To understand participants' views before taking part in the study, they were also asked to indicate how comfortable they might feel about donating their health data for the purposes of the databank outlined in the participant information sheet from 1 of the following categories: very comfortable, somewhat comfortable, not sure, somewhat uncomfortable, or very uncomfortable. Invitees were then sent a consent form by email, which they were asked to complete and sign. Patient and public members were offered a modest financial incentive for participating in the study in line with National Institute for Health Research guidance for recognizing public participation in research [24].

Focus Groups

Focus groups were conducted on the web on Zoom between March 24 and 31, 2022, and lasted for 90 minutes. A deliberative approach was used where focus groups began with a short presentation on the donated databank by a member of the research team, tailored to each stakeholder group, followed by a question-and-answer session so discussions could be fully informed. The proposed model presented to participants in the prediscussion presentations was of an opt-in approach where people would consent to donate their data to the databank. The facilitator did not direct discussions to confirm whether donated data would be identified or not so that participants could freely share their views around both scenarios.

The team employed a third-party organization with considerable experience in conducting focus groups on the topic of health data to facilitate the groups. Discussions were framed around 4 key questions: (1) Is having a donated free-text databank a good idea? (2) How best could the risks of holding donated, consented potentially identifiable data be managed? (3) What do you think about consent, and how it could be managed? and (4) Who should be allowed access, how should a databank be housed, and for what purposes? Multimedia Appendix 1 presents the questions asked of each stakeholder group.

Discussions were audio recorded, transcribed, and analyzed using thematic analysis.

Ethics Approval

The study was approved by the University College London REC (0976/002) and complies with the COREQ (Consolidated Criteria for Reporting Qualitative Research) [25] checklist for reporting qualitative studies.

Results

Overview

A total of 61 participants took part in the focus groups including patients and members of the public (24 participants), clinicians (10 participants), NHS Trust IG leads and REC members (14 participants), and NLP researchers (13 participants).

In total, 75% (46/61) of the participants recorded their demographic information on their expression of interest form. Of those, 54% (25/46) were female, 52% (24/46) were aged between 31 and 50 years, and 73% (33/45) were White. Overall, most participants (30/46, 66%) were either very comfortable or

somewhat comfortable donating their data to the databank, 28% (13/46) were not sure whether they would be willing to donate their data, and only 6% (3/46) of all participants felt either very or somewhat uncomfortable donating their data, all of whom were patients and public members, although the numbers were small (3 patients and public members out of 17; Table 1).

Key findings of the study are summarized in Multimedia Appendix 2.

Table 1. Participant demographic information and views around sharing their own data captured before participating in the focus groups.

	All participants (n=46[a]), n (%)	Patients and public members (n=17[a]), n (%)	Clinicians (n=10[a]), n (%)	Information governance leads and research ethics committee members (n=14[a]), n (%)	Natural language processing researchers (n=5[a]), n (%)
Sex					
Female	25 (54)	10 (59)	3 (30)	9 (64)	3 (60)
Male	21 (46)	7 (41)	7 (70)	5 (36)	2 (40)
Intersex	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Other	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Age group (years)					
≤30	3 (7)	2 (12)	0 (0)	0 (0)	1 (20)
31-50	24 (52)	4 (24)	8 (80)	8 (57)	4 (80)
51-65	12 (26)	6 (35)	2 (20)	4 (29)	0 (0)
>65	7 (15)	5 (29)	0 (0)	2 (14)	0 (0)
Ethnicity					
Asian or Asian British	7 (16)	3 (18)	3 (30)	0 (0)	1 (20)
Black, Black British, or African	2 (4)	2 (11)	0 (0)	0 (0)	0 (0)
Mixed or multiple ethnic groups	1 (2)	1 (6)	0 (0)	0 (0)	0 (0)
White	33 (73)	11 (65)	6 (60)	13 (100)	3 (60)
Other ethnic group	2 (5)	0 (0)	1 (10)	0 (0)	1 (20)
Views around donating data					
I would feel very comfortable donating my data	15 (33)	8 (47)	4 (40)	3 (21)	0 (0)
I would feel somewhat comfortable donating my data	15 (33)	5 (29)	3 (30)	6 (43)	1 (20)
Not sure	13 (28)	1 (6)	3 (30)	5 (36)	4 (80)
I would feel somewhat uncomfortable donating my data	2 (4)	2 (12)	0 (0)	0 (0)	0 (0)
I would feel very uncomfortable donating my data	1 (2)	1 (6)	0 (0)	0 (0)	0 (0)

[a]Data are based on 75% (46/61) of the participants who returned this information. Within stakeholder groups, data were returned as follows: patients or public members, 71% (17/24); clinicians, 100% (10/10); information governance leads or research ethics committee members, 100% (14/14); and natural language processing researchers, 38% (5/13).

Is Having a Donated Free-Text Databank a Good Idea?

Perceived Benefits and Challenges

Participants were very enthusiastic about the databank and its intended purpose and saw great value in establishing a platform for development and testing NLP tools to improve their accuracy. Many participants across groups articulated the benefits of producing trustworthy tools to unlock the rich data available in free text, which extended beyond improving NLP methods, including expediting access to, and use of, NHS data by speeding up permissions; accelerating development of NLP tools; and improving health and care leading to better outcomes for patients. The NLP researcher group highlighted its potential value as a training resource to teach and onboard researchers and help familiarize them with free-text data. NLP researchers and clinician groups both welcomed the opportunity to access free-text data for a UK-based population, which would be more appropriate for developing NLP tools on UK health care data sources, moving away from a reliance on US-based data such as MIMIC III or the recently released MIMIC IV [17,18]:

> *I'd love to see it come to fruition. I think it would be an absolute gold mine.* [IG lead and REC member]

> *Don't let this racehorse designed by [a] committee become a camel, just get something out. I think*

> *anything is better than what's currently offered, which is nothing.* [NLP researcher]

Several participants in the patient and public group felt that increased access to patient data as a result of the databank may prompt clinicians to improve the quality of their free-text data entry, as they will be more conscious of its wider use:

> *Very much a great idea. So, the MIMIC dataset I've worked with a lot has been really transformative for clinical NLP research in the US. But the MIMIC dataset has some serious issues, in terms of the kinds of data that are included, the representativeness of the sample, and so on, and so forth. So having something that can be created, as a research-specific resource like this, and created with more intentionality, and more design, as to what should be going into it, I think is a really, really incredibly valuable thing to do.* [NLP researcher]

Despite strong support, participants in all groups advised that clarity around the purpose of the databank; how it will be used; and by whom, both now and in the future, will be essential to its success. Patient and public participants felt this should be made clear in the consent process. Many participants, but patients or public members in particular, expressed concerns around how inaccurate recording, the lack of up-to-date data, or subjective data based on a physician's own impressions may threaten the aims of the databank, citing frequent instances of errors in their own health records. IG leads and REC members and patients or public participants felt that the accuracy of data must be improved before people trust the outcomes of the databank, and patients or public participants were keen for easier access to their own EHR so they can amend inaccuracies or missing information. However, other participants did not feel data accuracy would be a key factor in the success of the databank, and NLP researchers suggested the databank could provide a unique opportunity for investigating the effect of inaccurate or subjective training data on NLP research findings.

Biases owing to missing data or the lack of generalizability was a considerable concern among all groups. Many felt that donations will be more likely to come from White, middle-class populations and less likely to include people with rare diseases or whose records contain sensitive information. This was seen as most likely to affect data in mental health and social care settings:

> *There may be intrinsic biases in the actual data that's getting selected because certain classes of patients, whether it's by demographics, such as race or income, have more trust in what this is trying to achieve and, therefore, people with less trust won't actually consent to their data being used and we know, for instance, that that can be quite heavily in race, in the UK, on health data, for instance, and health services and the provision of health services. So, that may introduce biases in the data set.* [NLP researcher]

> *I'm not saying it would be necessarily unprofessional, but there could be things that may have been written 10 years ago and that maybe wouldn't be written now. Is that going to affect your data sets? So, I guess it's really about the bias that might be there within the unstructured data and whether you're proliferating that bias by collecting them and then training algorithms.* [NLP researcher]

Many participants saw artificial intelligence (AI) as playing a key role in health care in the future but were concerned about how AI tools are developed and perform in general. These concerns, which included questions around how tools "decide" which words to analyze, the possibility of scan reports being misread or missing key data, and the accuracy of annotation, are key to feeding into the communication plans for the databank.

Data Privacy and Use

Fears around data privacy for both patients and clinicians were raised. Participants discussed complexities in relation to free text, which might act as a barrier to data donation, and how clinicians may be uncomfortable that their identity and views are shared with researchers, for example, where GPs' personal views around a patient's health are recorded. Participants advised that such fears might be mitigated to some extent by ensuring robust data security, governance, and transparency around the "data pipeline"—that is, what the databank will be used for and by whom, for example, whether there is any commercial benefit. IG leads and REC members and patients or public participants in particular discussed challenges in articulating how tools may be used in the future and by whom, as technology evolves and society's views around acceptable and ethical use of their data may change over time. Participants thought that building scenarios for future use into the consent process is therefore important:

> *My only worry or concerns would be the way that technology develops way into the future and therefore, algorithms, as a result. And it could be ethical now but maybe it would be less ethical in the future.* [Patient or public member]

Types of Data to Include in the Databank and Data Linkages

All participants felt that it is very important that the databank benefit from the inclusion of data from a range of sources to reflect the whole systems approach of the health service and a more integrated care system of the future.

> *I don't see how, at this stage, you can start to select what you want to look at because you don't know what you want to look at. If the purpose of this is to develop algorithms for extracting useful, contextual information then you want the original data, which is being used to train the algorithm, to be as broad as possible.* [IG lead and REC member]

> *We are talking about a holistic approach, so that vision should be, I totally agree, whole patient records. And if we are going to use the computers, and train the algorithm, the whole purpose is looking at the wider picture, and bringing it together.* [Clinician]

Along with primary and secondary care data, some saw value in training NLP tools based on sources that are less commonly used for research, including social services and housing data, to encourage more research in these fields. Perceived advantages of including a broad range of data include introducing a more holistic view of a person's health needs and including training data based on the different styles of free-text data held in different settings. Some participants discussed the advantages of introducing the databank in phases to ensure timely access and build trust, for example, by starting with 1 health condition such as diabetes or mental health or geographical location.

Such phasing would need to be carefully considered. If condition-specific phasing were to be used, then the imperfections of the diagnostic codes used to select data for a phase would need to be addressed. Solutions might be to use existing NLP applications to identify conditions, accepting that this would have its own limitations, or choosing broader categories of data such as data from specific medical specialties or units:

> *Our health is based in our experience, and what starts in primary, might end up in secondary. You can't divide them into two separate things.* [Patient or public member]

There were mixed views around linking free-text data to other forms of data. NLP researchers particularly valued the opportunity to link the data to other sources, especially mental health data where rich narrative adds important additional detail for research, and some felt that the inclusion of coded data could help verify the accuracy of the free-text data more efficiently. However, some questioned the value of such linkages to train NLP tools and felt linkages, for example, to administrative data, may increase the risk of reidentification and would be resource intensive to manage:

> *In psychiatry, it tends to be that a lot of the information is locked behind this clinical free text, and they tend to be, by comparison to other medical specialties, a bit more verbose, a bit more narrative, in nature. And so, the ability to incorporate that, and the metadata that's required to have that, needs to be built in, I think, from the very, very start this has to be extensible, because I'd love to, for example, be able to look at GP notes, and see how they translate over into secondary care, but there are decisions about how this may be structured, early on, that could make that more difficult. So I think that's something that we need to build in from the start.* [NLP researcher]

IG leads and REC participants discussed the potential for linkages resulting in "scope drift," which could lead to the use of data outside of its original purpose. If this occurs, the databank should clearly communicate its wider remit. Scope drift was also a concern for linkages to patient-generated data such as wearable and monitoring devices. Some questioned the accuracy of these data sets and whether this may lead to the development of inaccurate NLP tools, although potential advantages including signaling support for greater use of patient data and "future-proofing" development of NLP tools by including these data, given the inevitable advancement and incorporation of these devices in health care, were also discussed:

> *This is where, as an IG person, I always start getting nervous and we're having conversations in our population health because everybody starts saying, "The police, it would be really good if we can put an algorithm together to identify individuals that might start being people that will cause domestic harm." So, that's an area that is always a bit nervous to say, "Well, what is that trying to establish," especially with this project. Because what's the purpose of having a linked dataset and then trying to do modelling on trying to extract free-text data? I'm not quite sure of the purpose of those two things together.* [IG lead and REC member]

Managing Risks of Holding Donated, Consented Identifiable Data

Data Privacy and Deidentification

The groups did not discuss a preference for whether data should be left in an identifiable form or should be deidentified; rather, discussions focused broadly on the issue of how to minimize the risk of reidentification, indicating that participants expected data to be deidentified. In particular, the patient or public group highlighted the importance of deidentification alongside data security, including robust data storage and management practices to mitigate risks to data privacy. All participants recognized that eliminating risk completely is unrealistic, but introducing steps to reduce the likelihood of reidentification by using birth year, partial postcode, or a sample of the notes rather than the whole record was discussed. Some clinician participants were conscious of potential legal implications for themselves and supported the removal of clinician details as well. Participants were mindful that the process of deidentification should be carried out on a case-by-case basis. NLP researcher members suggested involving data controllers in agreeing with the approach to deidentification, given their expert knowledge of the data and availability of resources such as data dictionaries to aid the deidentification process. Although rare diseases were regarded as posing a particular risk to reidentification, patients or public participants who themselves have been diagnosed with a rare condition were keen that this should not act as a barrier to the much-needed research and proposed that clear explanations about how data would be used, by whom, and what the data protection issues are might offset concerns. Alongside deidentification, it was felt that following the UK Caldicott Principles, which helps ensure confidential and appropriate use of people's data [26], adopting strong data security measures to protect against hacks and ensuring that only legitimate, vetted people have access to only the data they need were viewed as key to managing privacy concerns. Patients or public group members in particular voiced the importance of articulating these safeguards to reassure potential donors, and IG leads and REC members suggested that because deidentification is both difficult to define and achieve, there should be an emphasis on defining the purpose of accessing the data and the methods of safeguarding the data:

> *Is it a pseudonym? Is it a number? Is it aggregated? What level of anonymity is there when we're discussing this? What's proposed, or are there different levels (to de-id)?* [Patient or public member]

> *If you're doing a decent level of de-identification so you're getting masking rates of 90% or something, the risk is going to be very minimal, particularly if you're providing samples of notes rather than an entire record level. So if you're looking at medication as your concept for annotation, then you don't need stuff that doesn't contain that data, or is unlikely to. So you can start to pick how you pull your sample notes from a record. I think there's quite a lot you can do to continue to reduce and reduce and reduce that risk, but you won't eliminate it.* [IG lead and REC member]

Raising Awareness of the Databank

Future consultation with stakeholders, including clinicians, was viewed as essential by all groups, who were keen that engagement be continued to help develop the scope, consent model, and communication plan for the databank. Several participants advised careful planning on how to explain the databank. Patients or public participants expressed the importance of terminology when communicating plans, for example, to clarify what is meant by free text and what is in a health care record that might help allay people's fears around what they are agreeing to donate:

> *I want some power. I don't want to be a passive recipient of this whole data process, which is what's happened to a lot of us regarding data so far.* [Patient or public member]

Participants suggested that a targeted, small-scale approach be adopted in the early stages of raising awareness of the databank and starting to gather donations, working with trusted organizations who use health data for research. The involvement of GP practice staff in communicating the databank was seen as important, but clinician participants in particular were skeptical, saying that this was both impractical given their already stretched resources and unnecessary given patients already have the right to agree to their health records being shared. One alternative discussed was to recruit GPs or practice staff who were willing to take on this role. Some ideas for how to reach potential donors included making posters and leaflets available in GP surgeries with a QR code linked to a website about the databank, working with trusted organizations that support the use of health data for research such as *HealthWise* [27] and *use MY data* [28], and identifying community-based "public champions" to advocate the benefits and safety of the databank:

> *I would share it with patients but I'm constantly limited by time—something else to include in the consultation, so it has to be done through a different delivery mechanism than face to face.* [Clinician]

Managing Consent and Offering Choice

Participants expressed the importance of working with trusted NHS and research organizations and providing accessible consent including web-based consent, the availability of audio consent for people with low literacy levels, and translation into multiple languages. The information sheet should clearly set out the intended purpose of the databank and focus on providing reassurance regarding use and data security. If the approach was to include deidentified data, participants wanted this made clear in the supporting materials, possibly aided by including examples of "dummy" data to show what types of data are excluded in the deidentification process. Patients or public members thought that offering people the opportunity to view the databank before consenting might help them understand what information would be included:

> *People should know what they're getting involved in. Maybe seeing the database before they opt-in, and also, how far back will the data be taken from...* [Patient or public member]

> *Having an actual example of a fake letter or fake clinical notes with a lot of identifiable data and what, actually, will go to the databank in front of you so you see that they did remove all information about you talking about your kids or how your neighbour is annoying you because she is very noisy, all that kind of thing. Maybe, they would see it, that would make a lot of reassurance.* [NLP researcher]

Patients or public participants felt that offering choice over which data to donate and which type of organization can access the databank might increase donation rates, but others thought this may be complex and resource intensive to achieve and may encourage withholding of sensitive information. The IG leads and REC members group felt that offering choice with a promise to withhold sensitive information is likely to be unachievable and, therefore, undermine trust.

Who Should Be Allowed Access, How Should a Databank Be Housed, and for What Purposes?

Overview

Participants were in favor of access to the databank by both university- and NHS-based NLP researchers. Establishing a "road map" of the types of organizations with which the databank will work was suggested as helpful. Many participants favored a defined approach to access in the early stages of developing the databank, whereby access should be limited to NHS and university-based NLP researchers, and developing a set of standards could ensure that the use of the databank remains in line with its intended purpose. However, the IG leads and REC members group felt that it may be more appropriate to define the types of organizations that can access the databank based on a "compliance" model in which users should show that they can meet a defined level of capability and accountability. Standards should include an assessment of the applicant organization's motivation for using the databank and their reputation, although the road map should include a well-thought out vetting process to include applicants such as start-ups who may not have a proven track record in trustworthy access to data:

> *If there were particular requirements around use of the data, commitments to not attempt to re-identify,*

> *similar things like that, I think in my view both academic organisations, research students and also research organisations, with the right safeguards, even start-ups, I feel if they can meet a certain level of capability and accountability, then I feel that should be the bar rather than defining the type of organisation.* [IG lead and REC member]

Participants had mixed views on the use of the databank by commercial organizations including technology and pharmaceutical companies. Patients or public members were generally against the use of their data by these organizations for the purposes of the databank, although other groups felt that these partnerships may be beneficial owing to commercial organizations' considerable expertise and resources and that they should be allowed access if they could show that they meet the databank's standards. The primary consideration of commercial access was to ensure adequate return and public benefit. Participants discussed the need for assurances around how tools will be fed back into the public sector, for example, whether the NHS would have discounted access to any tools that were developed as a result of the databank. Access by charities was viewed as potentially problematic, as charities are less regulated and often have a campaign focus, which may result in data use being less scrutinized or controlled and the creation of NLP tools biased toward certain outcomes. Government organizations, insurance companies, and lawyers were deemed unsuitable.

Several participants agreed that existing models of good practice about access such as the Secure Anonymised Information Linkage (SAIL) Databank [29] in Wales should be incorporated and that learning from the Centre for Data Ethics and Innovation Public Attitudes Survey on Data and AI [30] about who is trusted with data and in what circumstances should inform the road map.

Fee-Based Model for Use of the Databank

Clinician and NLP researcher groups were asked about how the databank should be funded. Both groups welcomed government funding to help develop the databank, but a fee-based model was viewed as more sustainable in terms of supporting the management of access and oversight, ensuring data quality, and allowing data to be updated over time. Charging a not-for-profit fee was viewed as realistic, and participants favored the development of a tiered costing model with different levels of access and cost depending on the user, their reason for access, and the volume of data requested. Suggestions for tiered access included providing an institute-wide membership fee to enable access for anyone working within the organization and discounted rates or free access for those who contribute to data donation, maintenance of the databank, or data cleaning or other data quality control. It was suggested that the development of the cost model should be informed by existing approaches such as the Linguistic Data Consortium [31], which supports NLP research by creating and sharing resources, and the UK Data Service [32], a large repository of economic, social, and health data sets for research and teaching.

Data Gathering, Management, and Housing

Participants felt that with health care services under notable pressure, a robust and sustainable costing plan to support the transfer of data into the databank is essential. For example, GP practices may expect to be paid to carry out data extraction for the databank; therefore, costs should be factored into the costing model.

> *Services are already stretched and I know in some of my practices if you come to them and ask them for data, whether patient's consented or not, they're going to tell you to go take a high jump. I'm not going to spend my time extracting that for you or printing that off and sending it to you because I haven't got the time or capacity.* [Clinician]

The participants articulated 5 considerations that should be built into the way the databank is managed. The considerations were that the databank should be (1) accessible: access should be easier than the current process of applying for and accessing data from data providers; (2) up-to-date: the databank should be supplemented with new data so NLP tools are trained on current information; (3) controlled: robust technical controls should be put in place as the primary mechanism for managing data, including allowing and revoking access; (4) tracked: a mechanism should be in place to allow the use of data to be tracked in real time to see what happens to the data after they are accessed; and (5) transparent: transparency around how the databank is being used to track public benefit, for example, publishing details on a website, was seen as essential.

There was no clear consensus on the best approach to housing the databank. The options discussed included housing it within the NHS (perceived benefits were trustworthiness and existing robust policies around data breaches or data misuse; disadvantages were possible lack of technical infrastructure or resources to be able to manage it effectively) or a university setting. Different models considered included adopting a partnership approach in which the databank could be housed within a university but be governed by the NHS. Storing data within a secure environment such as a trusted research environment where data cannot be removed was felt to be appropriate, and existing models of good practice should be drawn upon, for example, Genomics England [33] and Health Data Research United Kingdom [19].

Oversight and Management

The integrity of the databank was closely linked to the approach to oversight and clear communication of how gatekeeping will take place. Most participants favored the establishment of an oversight committee to consider the range and types of data collected, review applications for access and use, ensure transparency around use including what the NLP tools being developed will be used for, and monitor and review data safeguarding. The committee should be independent and consist of people with diverse demographics, backgrounds, and expertise, including experts in the use of data, data controllers, and lay representatives. Patients or public participants discussed the importance of ensuring that the application process to join the oversight committee was accessible and did not put off

potential new applicants by overemphasizing a requirement for previous experience, as is often the case.

Discussion

Principal Findings

This study set out to test early stakeholder thinking around the acceptability and design considerations for the creation of a consented donated databank of clinical free text to develop and test NLP methods and tools. Understanding the details to inform establishment of such a databank was highlighted as a key recommendation in a recent position paper on the development of data governance standards for using clinical free-text data in health research [1]. All stakeholder groups voiced strong support and a pressing need for a free-text databank for the purposes set out to them. Participants highlighted a range of complex issues for consideration as the databank is developed, but there was a plea, particularly among the NLP researcher group, to move with haste to design something that works without becoming overburdened with the many complexities. One suggested approach was to develop the databank in phases, with the initial phase focusing on a specific health condition or type of health data, to test out whether people are willing to donate their data for this purpose and how it would work. Although not raised by any of the groups, a sensible starting point for the databank may be to exploit existing cohorts such as Generation Scotland or the UK Biobank where participants have already consented to share their data for research and can be easily contacted to invite them to donate their data to the databank. Participants stressed the importance of ongoing engagement and involvement with stakeholder communities in the development and operation of the databank. This position encapsulates the widely held view of the importance of transparency to increase the general lack of awareness about how patient data are used, by whom, and for what purposes [16].

Future proofing the databank at an early stage was viewed as important to take into account how uses of the databank and advances in technology might change over time. Participants also highlighted the importance of ensuring that plans for the provision and maintenance of the databank are sustainable in the future. There was a general agreement among stakeholder groups that the databank should draw upon a range of data sources to ensure that NLP tools reflect an integrated care system in the future, although there were mixed views about the benefits of linking to other data sources. The types of data to be included in the databank (eg, structured data in the GP or hospital record that may improve the performance of existing models owing to the inclusion of text-based features [34,35] or linkage to other data outside the NHS EHR, eg, national registry, mortality, or administrative data) should reflect stakeholder views on acceptability and practicalities, including cost, especially at the start. These issues should be explored in more detail in the next phase of the study. Participants agreed that the way the databank is created will be crucial to its success. The importance of communicating the intended purpose and approach to accessing the databank and the proposed mechanisms for safeguarding the data came up at numerous points during discussions with all groups. Learning from existing examples of good practice around access to data, data security, and how to fund the databank was also deemed crucial.

During the discussions, participants were not directed by the facilitator as to whether data stored in the databank would remain identifiable or not to encourage a broad conversation around anonymity in relation to the databank. Although it was made clear that donations would be based on explicit donor consent, issues of trust and maintaining patient confidentiality featured strongly in the discussions, particularly among the patient and public and IG and REC groups who were aware that public awareness of the risk of reidentification might act as a barrier to data donation, particularly among people with rare diseases who may be easier to identify. Interestingly, only patients and public participants expressed concern about sharing their own data with a databank (Table 1), although the numbers were small, with only 18% (3/17) of the patients or public members declaring that they would be very or somewhat uncomfortable donating their data. This finding is likely to reflect both a clearer understanding of the potential benefits of the databank among other stakeholder groups, given their expertise in this area, and the complexity of what is required to manage and mitigate the risks of data breaches and uphold privacy, which other stakeholder groups may understand more clearly. More work is needed to engage with patients and public members in this area to develop strategies for clear and widespread articulation of the benefits of the donated databank. Given the nature of the data and the challenges of removing personal identifiers, a realistic approach to deidentification will need to be adopted and made transparent and should be supported by robust processes to protect against risks. The deidentification approach may need to be dynamic, depending on the type of data and health condition, as some identifying data might be essential to the research study, for example, if the databank will be used to develop and test deidentification tools. It will also be worth exploring options to replace identifiers with random replacement identifiers to enable this type of work and remain mindful of the advances in generating synthetic data. Further exploration of stakeholder views to understand if stakeholders expect data to be stored in an identifiable form or if they want it to be deidentified is warranted. Although stringent steps will be adopted to minimize the risk of reidentification of patients by deidentifying the data and ensuring strict controls over who can access the data and how data will be made available, for example, within a trusted research environment, risk cannot be eliminated completely. The model for the databank should balance strong governance and security measures to ensure that access is not unnecessarily burdensome or complex. Learning from existing models will be the focus of the next phase of the development of the databank.

Another clear theme throughout the discussions was the need to develop a carefully planned and strong communication plan to build trust. It was suggested that distinct key messages be prepared depending on the stakeholders' interests in the databank. For example, there should be targeted communication with clinicians regarding their data privacy and with data controllers regarding data security. Communication should

incorporate relevant background information (eg, to counter the lack of awareness of what data are contained within EHRs) and address the context for the databank, which drives donations and the involvement of GP practices and other data providers.

Ideally, linguistic features in NLP should be representative of the entire population to ensure that the findings are not biased and are representative across patient groups. However, it is also important that the databank reflects what happens in the real world, despite the potential limitations owing to bias. The participants discussed the potential for bias and its impact on the databank ("garbage in, garbage out") in 2 areas: first, potential biases in the data because of inaccuracies or missing data, and second, potential biases in the data because of donations that lacked demographic variation. Biases in the annotation process [36] and other potential biases have not been discussed. Further in-depth consideration of how to avoid biases and the potential consequences for the trained models is needed when developing the databank and should be addressed clearly in the communication plan.

Participants expressed mixed views about the impact of training NLP models on inaccurate or missing data or data that are not up-to-date. Patients or public participants were concerned that the quality of data would affect the quality of the algorithms that will be developed. Patients or public participants were able to highlight numerous examples of inaccurate or missing data in their own health records which they found concerning and they expressed concern around how the potential lack of accurate or up-to-date data might impact on trust in the databank. Additional engagement with stakeholders, in particular patients and public participants, should be carried out to tease out and address major questions or lack of understanding about the impact of accuracy, subjectivity, and representativeness of data when training NLP tools so that people have a better understanding of what the databank can achieve. Communication should include efforts to make clear that NLP development is not interested in whether data are accurate or true, as it is simply trying to learn the linguistic properties of the data and how they relate to target concepts defined by NLP researchers and annotators.

Notably, participants were acutely aware of, and made reference throughout the discussions to, the important role AI will play in health care in the future but raised concerns about how AI tools are developed and perform and what might be ethical in the future. Embedding ethical approaches in developing data-driven technologies for AI and understanding public trust is high on the UK governments', researchers', and other stakeholders' agendas. For example, the NHS England Accelerated Access Collaborative [37] is committed to working with patients to ensure that AI innovations reflect the priorities of the end users and support innovators to embed public involvement in the development of AI technologies. The Centre for Data Ethics and Innovation, which is responsible for monitoring public attitudes toward data and AI over time, recently published findings from its second "Public Attitudes to Data and AI (PADAI) Tracker Survey" [30]. Findings from our study reflect similar views to those identified in this survey: for example, data security and privacy remain major concerns, people expect strong governance overseen by experts, and trust is strongly linked to the level of trust in the organizations that are accessing the data. Adhering to best practices around ethical AI principles and frameworks and anchoring public involvement in the development of the databank should be a priority to build trust, and developers of the databank should engage with leaders in the field to ensure this is embedded in plans for the databank, for example, the NHS England AI Ethics Initiative [38]. To keep up with the research and development in AI applied to clinical settings that is happening in the United States (made possible by data sets such as MIMIC III and IV), the UK government should channel resources into funding such a databank to harness rapid advances of AI technology and support long-term investment in the AI ecosystem in the future.

Limitations

This study has limitations. The sample size was relatively small, and there was a lack of diversity, particularly from younger and older participants and people of color. Therefore, the participants' views are unlikely to be representative of the UK general population. Engagement with more diverse groups and stakeholders who were not included in this work, for example, data controllers, is essential when planning next steps for the databank in the future. Thematic analysis does not allow views to be quantified, so we were unable to report how many participants felt a particular way. Furthermore, the aim of the focus groups was not to produce a specific set of recommendations. Rather, our findings provide useful insights into initial thinking, and the recommendations presented in this study (Textbox 1) therefore reflect a set of potential suggestions and advice based on the views of participants generally. Web-based discussions were limited to 90 minutes to ensure that the length of the focus groups was manageable for participants, which meant some topics could not be explored in depth. The research team therefore agreed in advance how to limit questions to ensure topics deemed the most relevant to particular stakeholders were covered within the time frame. Opportunities to explore topics in more detail will be sought in the next phase of the study.

Textbox 1. Proposed recommendations and suggestions for setting up the databank.

General approach

- Recommendation 1: Stakeholders should be involved throughout the development, implementation, and maintenance of the databank, including development of the scope, consent model, and communication plan.
- Recommendation 2: The databank should draw on the existing successful examples that can offer helpful models for consent, governance, data housing, and data security and be governed by an oversight committee.

Scope and phasing of the databank

- Recommendation 3: The databank should have a clearly defined purpose and take into account how natural language processing (NLP) researchers may wish to use it in the future.
- Recommendation 4: Development of the databank should be based on a small-scale, gradual approach to starting to gather donations to establish proof of concept and interest in donating to the databank. This might involve gathering data for 1 health condition (eg, diabetes) or location (eg, a mental health National Health Service [NHS] Trust) before moving to include others.

Channels to reach potential donors

- Recommendation 5: Reaching potential donors and publicizing the databank should include trusted individuals, networks, and organizations that support research using health data.
- Recommendation 6: Innovative ways to reach out to minority groups such as identifying public community champions who can advise and reassure others about the benefits and safety of the databank should be explored.

Consent

- Recommendation 7: Ensure the consent process is simple and accessible. Consent should be collected electronically, and information should link to a relevant NHS research ethics committee website and be offered in multiple formats and languages.
- Recommendation 8: The focus of the consent information sheet should be to provide reassurance around use; data security; and, if appropriate, deidentification. It should clearly define the purpose of the databank, provide a clear explanation of what data people are being asked to donate, and describe examples of scenarios for future use.
- Recommendation 9: Opportunities for showing potential participants their own personal health record before consenting should be explored.

Communication

- Recommendation 10: A clear and comprehensive communication plan should be carefully planned and developed with targeted messages for the different stakeholders (eg, clinicians regarding their data privacy and data controllers regarding data security).
- Recommendation 11: Communication should cover the following key elements clearly to build trust in the databank: predonation involvement (eg, possibility for participants to see their personal data and amend errors before donating); general aspects around data (what is free text? and what data are in a health care record?); content (what data are to be donated?); purpose (what the free-text data will be used for and by whom?); different contexts that NLP tools will be used in (eg, will data be used largely for commercial benefit?); and, crucially, the public benefit that NLP tools trained on the data could bring.

Pathways to databank access

- Recommendation 12: The foremost consideration for access should be to ensure public benefit and that benefits of data use are shared equitably.
- Recommendation 13: A “road map” should be developed to include the types of organizations the databank will work with, based on a compliance model where users should show they can meet a defined level of capability and accountability. The road map should include a set of standards and approach to “due diligence” to ensure databank use is in line with its intended purpose. Access could be granted based on an organization’s ability to meet, and commitment to comply with, the standards and an assessment of the applicant organization’s reputation and motivation for using the databank, rather than limiting which types of organization should be allowed access. The road map should incorporate recent learning on who is trusted with data and in what circumstances.
- Recommendation 14: Although development of the databank is in its infancy, it may be prudent to limit access to a small group such as NLP researchers linked to the NHS and UK universities.
- Recommendation 15: Access should be easier than the current process of applying for and accessing data from data providers.

Cost model for the databank

- Recommendation 16: A clear, transparent, and not-for-profit fee-based model should be developed that ensures sustainability of the databank over time. Fees should be used to maintain the database, support and manage access and standards, support oversight, ensure quality of data, and support updating the databank with new data over time.
- Recommendation 17: A tiered access model should include different levels of fees depending on the user, reason for use, and volume of data required (eg, access to a portion of the data set or all of it). Discounted or free access should be considered, for example, discounted access for organizations that contribute to data donation or where the databank would be used for teaching purposes. “In kind” arrangements could be considered for organizations that collaborate on improving quality of data (eg, cleaning data for access).

- Recommendation 18: Data provider costs to extract the data for the databank should be factored into the model.

Range and types of data to include in the databank

- Recommendation 19: A range of data types across different settings should be included in the databank.
- Recommendation 20: The databank should be kept up-to-date so that NLP tools are trained on current information.

Governance and oversight

- Recommendation 21: An independent oversight committee that has no stake in the NLP tools being developed should be established to monitor and review applications for data use and ensure transparency around use (eg, what the NLP tools being developed will be used for, the range and types of data included in the databank, who is using the databank, and safeguards to protect the data).
- Recommendation 22: The oversight committee should include members from a range of diverse ethnic and sociodemographic backgrounds and expertise, including data experts, data controllers, and lay representatives. Experts in specific data domains may be brought into the committee on an ad hoc basis to advise on specific applications.
- Recommendation 23: Ensure the process for applying to join the oversight committee is accessible and does overemphasize the need for previous experience on such committees.
- Recommendation 24: Lay members should be paid for their time serving on the committee.

Databank housing, management, and security

- Recommendation 25: The databank should be housed in a university or NHS trusted research environment.
- Recommendation 26: Robust technical controls should be put in place as the primary mechanism for managing data, including allowing and revoking access.
- Recommendation 27: A mechanism should be put in place to allow the use of data to be tracked in real time to see what happens to the data after they are accessed.
- Recommendation 28: Transparency around how the databank is being used to track public benefit should be ensured, for example, publishing details on a website.

Approaches to maintain participant anonymity

- Recommendation 29: Further engagement with stakeholders should be carried out to explore views around whether data should be stored in an identifiable or deidentified form in the databank and expectations around deidentification.

Comparison With Prior Work

Although the focus of this study centered on creation of the databank that has not, to the authors' knowledge, been previously explored, there were several areas where themes overlapped with previous research on attitudes toward the use of free-text data for research, which has been discussed particularly among patients and the public [16,39]. Although several benefits highlighted by participants in this study related specifically to the databank, wider benefits discussed included the potential for improving health and care leading to better outcomes for patients, which mirrored benefits identified by other UK research studies that used clinical free text [2]. Despite acknowledging a broad range of potential benefits, participants raised a number of concerns, particularly around how AI tools are developed and perform in general, the effect of possible biases, privacy risks, and reidentification. Previous research on potential harms of the use of free-text data for research has shown that the public harbors similar concerns around the use of free-text data for research generally, despite no evidence of these harms actually taking place following data breaches [40]. The issue of trust was raised several times, as was the importance of clear communication and a transparent approach to help build trust. Participants in this study felt that trust is strongly linked to the level of trust in the organizations that access the data, which echoes findings from other studies that showed that the public evaluates the trustworthiness of research organizations by assessing their competence in data handling and motivation for accessing the data [41]. A Citizens' Jury on the use of free text for research carried out in 2018 [16] found a high degree of willingness to share EHR data for public benefit among public participants who were informed about the use of free-text data, although participants expressed caution owing to concerns around the lack of transparency in the use of data and increased privacy risks. Participants in the Citizens' Jury suggested keeping patients informed about the use of their data and being transparent about ways to opt out of data sharing. These attitudes were mirrored in this study, as were views on risks related to deidentification of free text, which were in line with previous findings, including concerns around accuracy of removing patient identifiers.

Next Steps

The recommendations and advice resulting from the study are summarized in Textbox 1. The findings will be used to plan the next phase of developing the databank, including a pilot study to design the road map and communication plan and test the feasibility of donating to the databank. Next steps will involve identifying and reaching out to a broad range of stakeholders based on their diverse knowledge and skill sets to develop the vision for the databank and inform the road map and standards, including researchers, patients and public members, governance

experts, providers of NHS, data controllers, charities, government, and industry. The road map and standards could be further informed by a national web-based survey that will be co-designed with stakeholders to explore in more detail the acceptability and design considerations highlighted in this study, including understanding whether stakeholders expect data to be stored in an identifiable or deidentifiable form. Planning the next steps will draw on recommendations across relevant themes that were highlighted in a position paper on developing data governance standards for the use of free-text data in health research, including the involvement of patients and public members at identifiable data stages and opt-in consent models for the reuse of free-text data [1].

Conclusions

Improved access to clinical free-text data will help support technological innovation for developing novel and valid NLP tools to support research for public benefit. One way to leverage access is through the creation of a consented databank to develop and train NLP tools outside the NHS via the lawful basis of informed consent. This study showed strong multistakeholder support for a databank for this purpose and an urgent need to move forward to develop something quickly. Stakeholders expressed commonality around many issues such as governance, communication, and sustainability, but there were also stakeholder-specific concerns such as clinician concern around increased workload and privacy and patient-and-public concern around inaccuracies in their personal EHRs and how their data will be used. These issues should be explored in more detail and targeted among individual stakeholder groups. Findings from this study will be used to inform the next steps for adopting a partnership approach to establish a national, funded databank of free text for use by the research community.

Acknowledgments

This work was funded by Healtex and Health Data Research UK. The funders have no role in developing the content of this manuscript. The authors would like to acknowledge and thank all the focus group participants for their expert knowledge and continued support of this work and Hopkins Van Mil for facilitating the focus groups. ADS is supported by NIHR (AI_AWARD01864 and COV-LT-0009), UKRI (Horizon Europe Guarantee for DataTools4Heart) and British Heart Foundation Accelerator Award (AA/18/6/24223).

Data Availability

Focus group discussions were recorded for the purposes of report writing. Recordings were destroyed after being transcribed, but transcriptions are stored securely for 24 months, after which they will be destroyed (May 2024).

Authors' Contributions

All authors were involved in the conception and design of the study, critically reviewed the manuscript, and granted approval of the final version to be published. NKF led the participant recruitment and acquisition of data and wrote the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Topic guide: focus group questions and the distribution of questions between groups.
[DOC File , 88 KB - medinform_v11i1e45534_app1.doc]

Multimedia Appendix 2
Summary of focus group findings and key areas of discussion.
[DOCX File , 30 KB - medinform_v11i1e45534_app2.docx]

References

1. Jones KH, Ford EM, Lea N, Griffiths LJ, Hassan L, Heys S, et al. Toward the development of data governance standards for using clinical free-text data in health research: position paper. J Med Internet Res 2020 Jun 29;22(6):e16760 [FREE Full text] [doi: 10.2196/16760] [Medline: 32597785]
2. Ford E, Curlewis K, Squires E, Griffiths LJ, Stewart R, Jones KH. The potential of research drawing on clinical free text to bring benefits to patients in the United Kingdom: a systematic review of the literature. Front Digit Health 2021;3:606599 [FREE Full text] [doi: 10.3389/fdgth.2021.606599] [Medline: 34713089]
3. Dong H, Falis M, Whiteley W, Alex B, Matterson J, Ji S, et al. Automated clinical coding: what, why, and where we are? NPJ Digit Med 2022 Oct 22;5(1):159 [FREE Full text] [doi: 10.1038/s41746-022-00705-7] [Medline: 36273236]
4. ICD-10 Classification of Mental and Behavioural Disorders (The) Diagnostic Criteria for Research. Geneva: World Health Organization; 1993.

5. SNOMED CT (Systematized Nomenclature of Medicine -- Clinical Terms). Tech Target. URL: https://www.techtarget.com/searchhealthit/definition/SNOMED-CT [accessed 2022-12-05]
6. Poulos J, Zhu L, Shah AD. Data gaps in electronic health record (EHR) systems: an audit of problem list completeness during the COVID-19 pandemic. Int J Med Inform 2021 Jun;150:104452 [FREE Full text] [doi: 10.1016/j.ijmedinf.2021.104452] [Medline: 33864979]
7. CogStack homepage. CogStack. URL: https://cogstack.org/ [accessed 2022-12-05]
8. University of Dundee Health Informatics Centre homepage. University of Dundee. URL: https://www.dundee.ac.uk/facilities/health-informatics-centre [accessed 2023-02-24]
9. Data. Dataloch. URL: https://dataloch.org/data [accessed 2023-02-24]
10. Electronic Data Research and Innovation Service. Public Health Scotland. URL: https://www.isdscotland.org/products-and-services/edris/ [accessed 2023-02-24]
11. Clinical natural language processing research group. The University of Edinburgh. URL: https://www.ed.ac.uk/usher/advanced-care-research-centre/about/partners/clinical-natural-language-processing [accessed 2023-02-24]
12. Nind T, Sutherland J, McAllister G, Hardy D, Hume A, MacLeod R, et al. An extensible big data software architecture managing a research resource of real-world clinical radiology data linked to other health data from the whole Scottish population. Gigascience 2020 Sep 29;9(10):giaa095 [FREE Full text] [doi: 10.1093/gigascience/giaa095] [Medline: 32990744]
13. Dehghan A, Kovacevic A, Karystianis G, Keane JA, Nenadic G. Learning to identify Protected Health Information by integrating knowledge- and data-driven algorithms: a case study on psychiatric evaluation notes. J Biomed Inform 2017 Nov;75S:S28-S33 [FREE Full text] [doi: 10.1016/j.jbi.2017.06.005] [Medline: 28602908]
14. Public task. Information Commissioner's Office. URL: https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/lawful-basis-for-processing/public-task/ [accessed 2022-12-15]
15. General data protection regulation homepage. General Data Protection Regulation. URL: https://gdpr-info.eu/ [accessed 2022-12-05]
16. Ford E, Oswald M, Hassan L, Bozentko K, Nenadic G, Cassell J. Should free-text data in electronic medical records be shared for research? A citizens' jury study in the UK. J Med Ethics 2020 Jun 26;46(6):367-377 [FREE Full text] [doi: 10.1136/medethics-2019-105472] [Medline: 32457202]
17. Johnson A, Pollard TJ, Shen L, Lehman LW, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016 May 24;3:160035 [FREE Full text] [doi: 10.1038/sdata.2016.35] [Medline: 27219127]
18. Medical information mart for intensive care. MIMIC. URL: https://mimic.mit.edu/ [accessed 2023-02-24]
19. Health Data Research UK homepage. HRD UK. URL: https://www.hdruk.ac.uk/ [accessed 2022-12-05]
20. People in research. National Institute for Health Research. URL: https://www.peopleinresearch.org/ [accessed 2022-12-05]
21. National data guardian. GOV UK. URL: https://www.gov.uk/government/organisations/national-data-guardian [accessed 2022-12-05]
22. NHS Health Research Authority homepage. Health Research Authority. URL: https://www.hra.nhs.uk/about-us/committees-and-services/res-and-recs/search-research-ethics-committees/ [accessed 2022-12-05]
23. Healtex homepage. Healtex. URL: http://healtex.org/ [accessed 2022-12-05]
24. Payment guidance for researchers and professionals. National Institute for Health and Care Research. URL: https://www.nihr.ac.uk/documents/payment-guidance-for-researchers-and-professionals/27392 [accessed 2022-12-15]
25. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357 [FREE Full text] [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]
26. The caldicott principles. National Data Guardian. 2020. URL: https://www.gov.uk/government/publications/the-caldicott-principles [accessed 2022-12-05]
27. Health wise data homepage. HealthWise Data. URL: https://www.healthwisedata.com/ [accessed 2022-12-05]
28. Use my data homepage. Use My Data. URL: https://www.usemydata.org/ [accessed 2022-12-05]
29. The secure anonymised information linkage homepage. SAIL Databank. URL: https://saildatabank.com/ [accessed 2022-12-05]
30. Public attitudes to data and AI: tracker survey. Centre for Data Ethics and Innovation. 2022. URL: https://www.gov.uk/government/publications/public-attitudes-to-data-and-ai-tracker-survey [accessed 2022-12-05]
31. Linguistic data consortium homepage. Linguistic Data Consortium. URL: https://www.ldc.upenn.edu/ [accessed 2022-12-05]
32. UK data service homepage. UK Data Service. URL: https://ukdataservice.ac.uk/ [accessed 2022-12-05]
33. Genomics England homepage. Genomics England. URL: https://www.genomicsengland.co.uk/research/research-environment [accessed 2022-12-05]
34. Arnaud E, Elbattah M, Gignon M, Dequen G. Deep learning to predict hospitalization at triage: integration of structured data and unstructured text. In: Proceedings of the IEEE International Conference on Big Data (Big Data). 2020 Presented at: IEEE International Conference on Big Data (Big Data); Dec 10-13, 2020; Atlanta, GA, USA. [doi: 10.1109/bigdata50022.2020.9378073]

35. Chen P, Chen L, Lin Y, Li G, Lai F, Lu C, et al. Predicting postoperative mortality with deep neural networks and natural language processing: model development and validation. JMIR Med Inform 2022 May 10;10(5):e38241 [FREE Full text] [doi: 10.2196/38241] [Medline: 35536634]
36. Hovy D, Prabhumoye S. Five sources of bias in natural language processing. Lang Linguist Compass 2021 Aug 20;15(8):e12432 [FREE Full text] [doi: 10.1111/lnc3.12432] [Medline: 35864931]
37. NHS accelerated access collaborative. NHS England. URL: https://www.england.nhs.uk/aac/ [accessed 2022-12-05]
38. The AI ethics initiative. NHS England. URL: https://transform.england.nhs.uk/ai-lab/ai-lab-programmes/ethics/ [accessed 2022-12-05]
39. Ford E, Stockdale J, Jackson R, Cassell J. For the greater good? Patient and public attitudes to use of medical free text data in research. Int J Population Data Sci 2017 Apr 18;1(1):229 [FREE Full text] [doi: 10.23889/ijpds.v1i1.249]
40. Understanding Patient Data homepage. Understanding Patient Data. URL: https://understandingpatientdata.org.uk/weighing-up-risks [accessed 2022-12-05]
41. Stockdale J, Cassell J, Ford E. "Giving something back": a systematic review and ethical enquiry into public views on the use of patient data for research in the United Kingdom and the Republic of Ireland. Wellcome Open Res 2018;3:6 [FREE Full text] [doi: 10.12688/wellcomeopenres.13531.2] [Medline: 30854470]

Abbreviations

AI: artificial intelligence
COREQ: Consolidated Criteria for Reporting Qualitative Research
EHR: electronic health record
GP: general practitioner
ICD-10: International Classification of Diseases, 10th revision
IG: information governance
MIMIC: Medical Information Mart for Intensive Care
NHS: National Health Service
NLP: natural language processing
REC: research ethics committee
SAIL: Secure Anonymised Information Linkage
SNOMED CT: Systematized Nomenclature of Medicine Clinical Terms

Edited by C Lovis; submitted 05.01.23; peer-reviewed by B Alex, S Machinathu Parambil Gangadharan, M Elbattah; comments to author 28.01.23; revised version received 24.02.23; accepted 19.03.23; published 03.05.23.

Please cite as:
Fitzpatrick NK, Dobson R, Roberts A, Jones K, Shah AD, Nenadic G, Ford E
Understanding Views Around the Creation of a Consented, Donated Databank of Clinical Free Text to Develop and Train Natural Language Processing Models for Research: Focus Group Interviews With Stakeholders
JMIR Med Inform 2023;11:e45534
URL: https://medinform.jmir.org/2023/1/e45534
doi:10.2196/45534
PMID:37133927

Original Paper

Acquisition of a Lexicon for Family History Information: Bidirectional Encoder Representations From Transformers–Assisted Sublanguage Analysis

Liwei Wang[1], MD, PhD; Huan He[1], PhD; Andrew Wen[1], MSc; Sungrim Moon[1], PhD; Sunyang Fu[1], PhD; Kevin J Peterson[2], PhD; Xuguang Ai[3], MSc; Sijia Liu[1], PhD; Ramakanth Kavuluru[4], PhD; Hongfang Liu[1], PhD

[1]Department of Artificial Intelligence and Informatics, Mayo Clinic, Rochester, MN, United States

[2]Center for Digital Health, Mayo Clinic, Rochester, MN, United States

[3]Department of Computer Science, University of Kentucky, Lexington, KY, United States

[4]Division of Biomedical Informatics, Department of Internal Medicine, University of Kentucky, Lexington, KY, United States

Corresponding Author:
Hongfang Liu, PhD
Department of Artificial Intelligence and Informatics
Mayo Clinic
200 First Street SW
Rochester, MN, 55905
United States
Phone: 1 507 293 0057
Email: liu.hongfang@mayo.edu

Abstract

Background: A patient's family history (FH) information significantly influences downstream clinical care. Despite this importance, there is no standardized method to capture FH information in electronic health records and a substantial portion of FH information is frequently embedded in clinical notes. This renders FH information difficult to use in downstream data analytics or clinical decision support applications. To address this issue, a natural language processing system capable of extracting and normalizing FH information can be used.

Objective: In this study, we aimed to construct an FH lexical resource for information extraction and normalization.

Methods: We exploited a transformer-based method to construct an FH lexical resource leveraging a corpus consisting of clinical notes generated as part of primary care. The usability of the lexicon was demonstrated through the development of a rule-based FH system that extracts FH entities and relations as specified in previous FH challenges. We also experimented with a deep learning–based FH system for FH information extraction. Previous FH challenge data sets were used for evaluation.

Results: The resulting lexicon contains 33,603 lexicon entries normalized to 6408 concept unique identifiers of the Unified Medical Language System and 15,126 codes of the Systematized Nomenclature of Medicine Clinical Terms, with an average number of 5.4 variants per concept. The performance evaluation demonstrated that the rule-based FH system achieved reasonable performance. The combination of the rule-based FH system with a state-of-the-art deep learning–based FH system can improve the recall of FH information evaluated using the BioCreative/N2C2 FH challenge data set, with the F1 score varied but comparable.

Conclusions: The resulting lexicon and rule-based FH system are freely available through the Open Health Natural Language Processing GitHub.

(JMIR Med Inform 2023;11:e48072) doi:10.2196/48072

KEYWORDS

electronic health record; natural language processing; family history; sublanguage analysis; rule-based system; deep learning

Introduction

Family history (FH) has long been regarded as a core element in caring for patients who have varied health concerns [1], with the capability to significantly enhance the delivery of precision medicine [2]. However, FH data are underused for actionable risk assessment [1]. One barrier to using FH information is provider preference in recording the collected FH information

in an unstructured format (eg, clinical notes) [3] as opposed to within electronic health record (EHR) structured data [4]. As clinical text tends to be unstructured, the information contained within is computationally inaccessible relative to that contained in structured records. This lack of computational accessibilities poses a challenge in using FH information for downstream data analytics or clinical practice (eg, via clinical decision support). One approach to render information data computationally accessible is through the use of natural language processing (NLP), thus motivating our work to develop an NLP system that can extract and normalize FH information.

Despite the majority of clinical NLP measurement studies focusing on statistical approaches, rule-based NLP systems based on semantic lexicons and rule patterns are popular among observational studies [5] for cancer research and practice [6,7]. With the advantages of ensuring process transparency, implementability, and scientific rigor, semantic lexicons and rule patterns are interpretable and easily modifiable, conforming to the FAIR (Findable, Accessible, Interoperable, and Reusable) and RITE (Reproducible, Implementable, Transparent, and Explainable) principles [8,9] for scientific data management. In addition, semantic lexicons and rule patterns capture sublanguage characteristics of domains that can be portable and generalizable to other applications [10]. By sublanguage, we refer to domain-specific linguistic and lexical patterns that are more prominent in free text in specialized fields such as medicine.

One popular lexical resource for clinical NLP is the Unified Medical Language System (UMLS), a repository of biomedical vocabularies distributed by the US National Library of Medicine, integrating over 200 biomedical vocabularies. A source vocabulary contained within the UMLS is the Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT), which is the recommended coding system for clinical problems. As not all terms in the UMLS or SNOMED-CT are part of the FH sublanguage, in this study, we exploited a corpus-driven method with pretrained language models to build an FH semantic lexicon with the normalization feature and reasonable size and coverage.

There have been previous efforts focused on creating semantic lexicons for clinical NLP. Johnson [11] automatically constructed a semantic lexicon based on the Specialist Lexicon of the UMLS, which can assist NLP analysis of a medical narrative with the semantic preference options of selecting semantic type. Luo et al [12] created a semantic lexicon using UMLS knowledge sources by leveraging a corpus from ClinicalTrials.gov. Liu et al [13] constructed a corpus-driven semantic lexicon based on the UMLS assisted by variants mined and usage information gathered from clinical text.

Regarding deep learning–based approaches, pretrained language models such as bidirectional encoder representations from transformers (BERT) [14] can learn the structure of language (ie, the basic semantic and syntax information) through unsupervised training on a large corpus of unlabeled text [15]. Given a new task, such pretrained models can be fine-tuned with a small number of annotated samples to perform well [16].

With respect to the FH information extraction task, we hypothesized that terms in a large-scale corpus having semantic types similar to the entities labeled in the data set used for fine tuning can be detected by fine-tuned BERT models in a named entity recognition task. Additionally, we hypothesized that corpus-driven methods would enable more term variants to be discovered from real-world EHR data, from which the lexicon results could further enhance and empower FH information extraction systems. Operating under these two hypotheses, we here present an FH lexicon derived through a combination of these two approaches. To demonstrate the usability of the FH lexicon, we further developed a rule-based FH system based on the lexicon that extracts FH entities and relations specified in previous FH challenges and evaluated its performance accordingly. In terms of system development, we consider that an FH system that prioritizes recall is highly desired for NLP-assisted curation in EHR-based studies.

Methods

FH Concepts

Before assembling an FH lexicon, we defined FH concepts as those belonging to the selected UMLS semantic types within the “DISO” (disorder) semantic group (Table 1), excluding T050 (experimental model of disease) [17]. Figure 1 shows our study design. We first fine-tuned Bio_ClinicalBERT, UmlsBERT, and bert-base-uncased models, and selected the model with the best performance. We then used the selected model to extract potential disorder/finding related mentions in a large clinical corpus. Subsequently, potential FH mentions were automatically normalized, manually curated, and prepared into symbolic lexicon format compatible with the Open Health Natural Language Processing (OHNLP) Toolkit’s NLP engine MedTagger [18]. A coverage evaluation was conducted for the lexicon. To demonstrate the usability of the lexicon, a rule-based FH system was developed to extract FH information. In parallel, we also experimented with a deep learning–based FH system for FH information extraction. Previous FH challenge data sets were used for evaluation [19,20].

Figure 1. Study design. BERT: bidirectional encoder representations from transformers; ECH: employee and community health; FH: family history; I2B2: 2012 i2b2 natural language processing challenge data set (training data set); PURE: Princeton University Relation Extraction; SFCP: standardization framework for clinical problems.

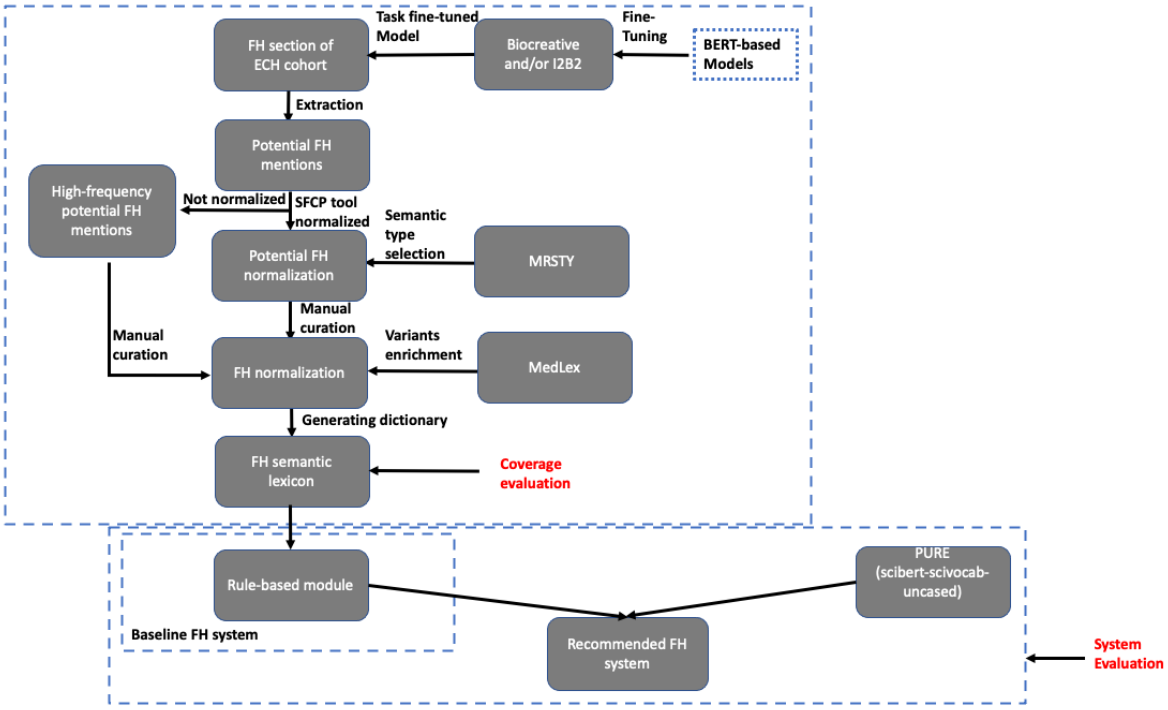


Table 1. Selected semantic types as family history concepts.

Semantic type code	Semantic type term
T019	Congenital abnormality
T020	Acquired abnormality
T037	Injury or poisoning
T047	Disease or syndrome
T048	Mental or behavioral dysfunction
T049	Cell or molecular dysfunction
T190	Anatomical abnormality
T191	Neoplastic process
T033	Finding
T046	Pathologic function
T184	Sign or symptom

Resources

Overview

Here, we introduce the resources used for lexicon construction and all relevant evaluations. Specifically, the 2018 BioCreative FH challenge training set and the 2012 I2B2 training set were used as the supervised data sets to fine-tune the deep learning models for lexicon preparation. Various data sets were experimented with based on a hypothesis that more data could encompass more semantic contexts of potential FH mentions. The 2018 test set was used to evaluate lexicon coverage. The 2018 and 2019 FH challenge training sets were used for training of the rule-based FH system and fine-tuning the deep learning–based FH system, and the 2018 and 2019 FH challenge test sets were used for evaluating the performance of the FH systems in extracting FH entities (task 1) and relations (task 2). A large EHR corpus was used for collecting potential FH mentions. MedLex was used to further enrich term variants of FH concepts extracted by the selected model. The UMLS was used for semantic type selection, while UMLS and SNOMED-CT were used for a comparison of size with our corpus-driven dictionary. Table 2 summarizes specific applications of the resources in the lexicon construction and all relevant evaluations. Detailed descriptions for each resource are provided below.

Table 2. Specific applications of resources in the lexicon construction and evaluations.

Application	A[a]	B[b]	C[c]	D[d]	E[e]	A+E	EHR[f] corpus	MedLex	UMLS[g]	SNOMED-CT[h]
Lexicon construction										
Fine-tuning the BERT[i] model	✓				✓	✓				
Dictionary entry collection							✓			
Dictionary concept enrichment								✓		
Semantic type selection									✓	
FH[j] system development										
Development of rule-based FH system	✓	✓								
Fine-tuning deep learning–based FH system	✓	✓								
Evaluation										
Dictionary coverage			✓						✓	✓
Challenge task 1			✓	✓						
Challenge task 2			✓	✓						

[a]A: Training set (BioCreative).

[b]B: Training set (N2C2).

[c]C: Testing set (BioCreative).

[d]D: Testing set (N2C2).

[e]E: Training set (I2B2/2010).

[f]EHR: electronic health record.

[g]UMLS: Unified Medical Language System.

[h]SNOMED-CT: Systematized Nomenclature of Medicine Clinical Terms.

[i]BERT: bidirectional encoder representations from transformers.

[j]FH: family history.

Synthetic FH Annotation Data Sets (A-D)

As the organizer of the BioCreative/OHNLP 2018 Family History Extraction Task [19] and 2019 NLP Clinical Challenge (N2C2)/OHNLP shared task [20], we curated deidentified annotation data sets based on synthetic clinical narratives. Data set A corresponds to data set C in the BioCreative Challenge and data set B corresponds to data set D in the N2C2 Challenge. FH was annotated as "observation" and defined as any health-related problem, including diseases, smoking, suicide, and drinking, while excluding auto accidents, surgeries, and medications [19]. Family members (FMs), observation, age, and living status were annotated as entities, and then all entities related to an FM category were linked into one chain. We further enhanced the resulting annotations in the data sets by normalizing observations to SNOMED-CT codes and correcting errors in previous annotations. The reannotated data sets are accessible based on the Data Use Agreement. Multimedia Appendix 1 shows the statistical comparison between original and enhanced annotations.

I2B2 Data Set (D)

The 2012 I2B2 NLP challenge organizers provided a fully deidentified data set with annotations for temporal relations as well as those generated from previous challenges such as the 2010 challenge of clinical concept extraction (problems, tests, treatments) [21], where problems include symptoms, complaints, diseases, and diagnoses.

EHR Raw Corpus

The raw corpus used in the study consists of 9,426,352 text segments extracted from the Family History section of clinical notes prior to 2013 of a primary care cohort (ie, the employee and community health cohort), which contains 83,000 patients at Mayo Clinic.

Dictionary Resources

MedLex is a semantic lexicon built on a large corpus of clinical documents collected at Mayo Clinic and from the UMLS (2011AA version) [13]. MedLex contains term variants from real-world EHRs, serving as a practical dictionary resource for FH lexicon expansion. We used MedLex to further enrich term variants of FH concepts extracted by the selected model. The UMLS (2021AA) and SNOMEDCT_US concepts accessible through the UMLS were restricted to only English entries. The MRCONSO table of the UMLS, which includes over 200 source vocabularies, was used for coverage evaluation, and the MRSTY table of the UMLS was used for screening semantic types for each concept unique identifier (CUI) in the MRCONSO table.

Ethical Approval

Use of the EHR raw corpus data was approved by the Mayo Clinic Institutional Review Board (17-003030) for Human Subject Research.

BERT-Based Corpus Analysis for Lexicon Construction

BERT Models for Extraction of Potential FH Mentions

BERT-base-uncased was pretrained on an unsupervised NLP data set using a masked language modeling approach [14]. Bio_ClinicalBERT was initialized from BioBERT and trained on all Medical Information Mart for Intensive Care notes [22]. UmlsBERT is a contextual embedding model that integrates domain knowledge during the pretraining process via a novel knowledge augmentation through the UMLS Metathesaurus [23]. UmlsBERT and Bio_ClinicalBERT are domains related to this study, while BERT-base-uncased could be used as a baseline comparison. We used configurations mostly consistent with the recommendations in the original release of the models. The maximum sequence length was set to 512, the batch size was set to 16, the total number of training epochs was set to 100, and the weight decay was set to 0.01. An early stopping method was used to determine the optimal number of epochs and to prevent overfitting. The train/test split was 80/20, where the "train" split was used for training and the "test" split was used for validation.

We fine-tuned these models on the BioCreative training data set, I2B2 training data set, as well as the combination of the BioCreative and I2B2 training data sets. We then selected the model with the best performance to extract any potential FH mentions in the raw corpus having coarse-grained semantic types similar to the entities labeled in the supervised data.

Normalization

The extracted FH mentions were automatically normalized through a standardization framework for clinical problems (SFCP) [24]. This framework converts free-text clinical problem descriptions into standardized forms based on the UMLS CUI corresponding to SNOMED-CT concepts and the Health Level 7 Fast Healthcare Interoperability Resources (FHIR)-based structured representations, including the codified problem and all relevant modifiers and context. The CUIs associated with the SNOMED-CT concept were used for coding. For example, for the mention "allergy-induced asthma," the framework outputs "C0440102 | Various patch test substance" and "C0155877 | Allergic asthma."

Manual Curation and Further Enrichment

We reviewed all normalized FH mentions and retained the main normalized problem concepts with semantic types corresponding to those previously selected. For example, for the mention "allergy-induced asthma," the codified problem "C0155877 | Allergic asthma" was kept and "C0440102 | Various patch test substance" was removed from the final lexicon. In some cases, one BERT-extracted mention can be normalized to several individual concepts through the automatic standardization. We then enriched the FH lexicon by keeping all individual concepts and obtaining associated variants from MedLex.

In addition, we manually mapped high-frequency mentions that occurred across at least 20 patients that were not automatically standardized to corresponding CUIs.

Rule-Based FH System

To demonstrate the capability of the FH lexicon in extracting FH relations, we further implemented a rule-based FH system by integrating rules for FM identification with the resulting lexicon. Multimedia Appendix 2 shows three degrees of consanguinity we aggregated into FM identification rules. FH relations were then extracted based on co-occurrence within a clause of one sentence or across three adjacent sentences if coreference existed, as indicated by keywords such as "he," "she," "none of them," "her," or "his," while excluding relations between FMs of spouse and FH. We implemented this as an OHNLP Toolkit module with code available on GitHub [25]. The implementation provides several output formats, including FHIR-based output, with FM and FH standardization conforming to FHIR standardization. The final output of the rule-based FH system includes entities and relations. The entity output includes file name (which links to document references with patients' ID), sentence ID, chunk ID, entity type, concept, and certainty. The relation output includes file name, FM, side of family, text of observation, and certainty. An option is also available to output this information to CSV, instead of FHIR, format. To set SNOMED-CT condition codes as the standard, a separate mapping file is required due to SNOMED-CT licensing restrictions.

Deep Learning–Based FH System

As information extraction remains a challenging task, it is preferred to investigate what is the gain from deep learning–based models. Therefore, we further implemented a deep learning–based FH system as follows. Note that we experimented with fined-tuned models for two purposes in this study. The first was to fine-tune models for identifying and collecting potential FH mentions from clinical texts to build a dictionary, as described in the previous section. Here, the second purpose was to fine-tune models for information extraction to automatically identify FH entities and relations.

The Princeton University Relation Extraction (PURE) system is an approach where the entity model builds on span-level representations and the relation model builds on contextual representations specific to a given pair of spans. As this pipelined approach has been demonstrated to be extremely effective, we implemented PURE using scibert-scivocab-uncased as the base encoder and fine-tuned it based on the BioCreative and N2C2 training data [26].

Evaluation

Overview

We conducted two evaluation studies, including (1) a coverage evaluation of the lexicon and (2) a comparison study of a lexicon-based module with a deep learning–based module for FH information extraction.

Lexicon Coverage

To the best of our knowledge, our lexicon is the first to incorporate a large number of text variants and concepts of FH. Therefore, we compared the resulting lexicon with the UMLS and SNOMED-CT in terms of the number of concepts under each semantic type. We analyzed the lexicon coverage by

calculating the number of concepts under each semantic type. In addition, we calculated the number of concepts and variants covered by the corpus-driven lexicon run against the BioCreative testing data set relative to annotated gold standards. True positive (TP) rate based on a partial match, false negative (FN) rate, and recall (TP/[TP+FN]) at the concept level and variant level were calculated.

Performance of FH Information Extraction

We evaluated the utility of the lexicon in identifying mentions of FMs and their associated attributes (side of family) using the BioCreative testing set and the N2C2 testing sets (ie, task 1 of the challenges). Precision, recall, and F1-scores were calculated as the performance metrics of the lexicon-based module, the deep learning–based module, and both.

We also evaluated the rule-based FH system's ability to identify relations between FMs, observations, and living status using the BioCreative testing set and N2C2 testing sets (ie, task 2 of the challenges). This task was different between the 2018 BioCreative and 2019 N2C2 challenges in that the latter added a certainty attribute (negated or nonnegated) into relation extraction. Three sets of precision, recall, and F1 values were separately calculated using varying setups: rule-based module only, deep learning–based module only, and a combination of both. Our evaluation scheme is the same as that applied in the 2018 BioCreative and 2019 N2C2 FH challenges. We performed a general error analysis to investigate error sources. In addition, to further investigate how much the lexicon contributes to system performance, we performed an ablation study and specific error analyses.

Results

Multimedia Appendix 3 shows the performance of the BERT-base-uncased, UmlsBERT, and Bio_clinicalBERT models fine-tuned on the BioCreative training data set, I2B2 training set, and on the combination of the BioCreative training set with the I2B2 training set. As the model fine-tuned on the combination of the two data sets outperformed other models, we selected the combined model to extract potential FH mentions from the corpus.

There were 72,518 unique entities identified by the Bio_clinicalBERT model fine-tuned on the combination data set, of which 47,250 (65.16%) were automatically normalized to 10,579 CUIs through the standardization framework for clinical problems. We manually normalized 148 entities occurring across more than 20 patients that were not automatically normalized to CUIs. For example, "typediabetes" with a frequency of 3693 was normalized to C0011854 (diabetes). Note that spellings such as "typediabetes" found in the EHRs are most likely typos due to physicians' writing, backend EHR note processing, or tokenization challenges. Therefore, manual normalization is important. After semantic type screening, manual curation, and MedLex enrichment, the final FH lexicon contained 33,351 dictionary entities normalized to 6177 CUIs and 15,126 SNOMED-CT codes, with an average of 5.4 variants for each concept. Table 3 shows the comparison of sizes of various lexicons. The corpus-driven lexicon was more light-weighted, with more variants per concept. This implies that implementation with the lexicon for NLP tasks would be easier and more efficient with the corpus-driven lexicon than with the SNOMED-CT and UMLS lexicons.

Lexicon coverage evaluation on the BioCreative testing data set showed that there are 137 TP entities corresponding to 128 concepts and there are 16 FN entities, of which 6 entities had no corresponding concepts in the FH lexicon and 10 entities had 10 corresponding concepts in the FH lexicon. Concept-level recall was 95.8% and variant-level recall was 89.5%. For the N2C2 testing set, there were 507 TP entities corresponding to 214 concepts and 62 FN entities, of which 33 entities had no corresponding concepts in the FH lexicon and 29 entities had 26 corresponding concepts in the FH lexicon. Concept-level recall was 87.9% and variant-level recall was 89.1%. Table 4 shows the comparison of numbers of semantic types of CUIs in various lexicons. It can be observed that the corpus-driven lexicon contains less concepts under each semantic type compared with the UMLS and SNOMED-CT lexicons.

Table 5 shows the performance of FH systems for subtasks 1 and 2 of the BioCreative and N2C2 challenge data sets (original and reannotated) categorized by the rule-based FH system only, deep learning–based FH system only, and a combination of the two. For task 1, the highest F1 score was 0.8766 from the deep learning–based model on the original BioCreative data set and was 0.8061 on the original N2C2 data set. For task 2, the highest F1 score was 0.6206 from the deep learning–based module on the original BioCreative dataset and was 0.5940 from the combined results on the original N2C2 data set. The rule-based FH system based on the corpus-driven lexicon produced lower F1 scores in contrast with the deep learning–based FH system for both tasks, but higher or comparable recall for task 1 and higher recall for task 2. The combined results had the highest recall compared with the rule-based FH system or the deep learning–based FH system for task 1, ranging from 0.8669 to 0.9475. The combined results also showed the highest recall (ranging from 0.7109 to 0.8370) and varied F1 scores (ranging from 0.4288 to 0.6142) for task 2.

Table 3. Comparison of the size of lexicons.

Lexicon	Concepts (CUI[a])	Variants	Average number of variants per CUI
Corpus-driven lexicon	6177	33,351	5.40
SNOMED-CT[b]	412,027	1,349,838	3.28
UMLS[c]	4,440,279	9,569,507	2.16

[a]CUI: concept unique identifier.

[b]SNOMED-CT: Systematized Nomenclature of Medicine Clinical Terms.

[c]UMLS: Unified Medical Language System.

Table 4. Statistical summary for semantic types of concept unique identifiers (CUIs) in the lexicon.

Semantic type code	Semantic type term	Number of CUIs		
		Corpus-driven lexicon	UMLS[a]	SNOMED-CT[b]
T019	Congenital abnormality	61	11,237	7034
T020	Acquired abnormality	62	4326	2105
T037	Injury or poisoning	236	113,258	29,371
T047	Disease or syndrome	2645	113,780	45,946
T048	Mental or behavioral dysfunction	376	9152	3318
T049	Cell or molecular dysfunction	24	4175	583
T190	Anatomical abnormality	94	6943	1703
T191	Neoplastic process	866	43,532	11,054
T033	Finding	1018	309,971	44,547
T046	Pathologic function	403	27,601	9098
T184	Sign or symptom	389	14,167	2931

[a]UMLS: Unified Medical Language System.

[b]SNOMED-CT: Systematized Nomenclature of Medicine Clinical Terms.

Table 5. Evaluation results for family history (FH) information extraction.

Task and data set	Deep learning–based FH system			Rule-based FH system			Combined		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Task 1 Original									
2018 testing	0.8819	0.8729	0.8766	0.7877	0.8619	0.8211	0.7857	0.9475	0.8469
2019 testing	0.8271	0.7835	0.8061	0.7191	0.8408	0.7740	0.7104	0.9007	0.7835
Task 1 Reannotation									
2018 testing	0.8830	0.8607	0.8709	0.7988	0.8294	0.8125	0.8081	0.9193	0.8486
2019 testing	0.7860	0.7747	0.7806	0.7149	0.7980	0.7536	0.7041	0.8669	0.7643
Task 2 Original									
2018 testing	0.7189	0.5464	0.6206	0.5413	0.6000	0.5673	0.5518	0.7237	0.6142
2019 testing	0.6841	0.5109	0.5851	0.4089	0.5571	0.4716	0.4699	0.8370	0.5940
Task 2 Reannotation									
2018 testing	0.6777	0.5307	0.5962	0.5596	0.5980	0.5769	0.5573	0.7109	0.6142
2019 testing	0.3309	0.8168	0.4629	0.4003	0.5292	0.4548	0.3060	0.8150	0.4288

A general error analysis showed that errors could be divided into two major sources. First, varied definitions of FH between different subtasks represented a confounding factor. FH in the gold standards was defined as any health-related problem, including diseases, smoking, suicide, and drinking, excluding auto accident, surgery, and medications [19], whereas the definition in the FH lexicon is based on semantic types (Table 1). For example, mastectomy, a procedure, was annotated as an

observation (FH) in the challenge data sets, which is not in the scope of our lexicon. Second, the intrinsic difficulty of FH relation extraction and its textual representations presented several obstacles, particularly with symbolic systems. The rule-based FH system used simple heuristic rules, and therefore it is difficult to handle complex relationships, especially when the subtask 2 of challenges includes multiple layers of relationships.

In the ablation study, for task 2 based on the original BioCreative test set, the precision, recall, and F1 score for observation only (excluding living status) were 0.5419, 0.6265, and 0.5783, respectively, representing a slight improvement compared with the corresponding values of 0.5413, 0.6000, and 0.5673 for both observation and living status. For task 2 based on the original N2C2 test set, the precision, recall, and F1 score were 0.3984, 0.7033, and 0.5449, respectively, for observation only (excluding living status); 0.4275, 0.6658, and 0.5444, respectively, for both observation and living status; and 0.3730, 0.6675, and 0.5169, respectively, for both observation and certainty. Although living status and certainty alone had little impact on the performance, the combination of observation, living status, and certainty resulted in significantly lower performance of 0.4089, 0.5571, and 0.4716 for precision, recall, and F1, respectively.

Discussion

FH has its own sublanguage. However, some terms related to FH may not actually be used in practice or may be used very rarely, such as “ancestor,” “descendant,” or “genealogy.” For this reason, these terms may not appear in the lexicon. As FH is specifically related to blood relations (consanguinity), it relates to the patient themselves. Therefore, the FM of spouse should not be considered, and FH elements relating to a spouse (rather than the patient) should consequently not be extracted. The advantage of this operation is the consistency with definitions of FM, resulting in a list of FH of the relevant FMs. The disadvantage may be missing the spouse’s relatives and associated FH information. There are some social, behavioral, and environment factors shared in the same household, which also represent critical information. However, these are social determinants of health and not part of the blood relations.

Collecting lexicon entries can be defined as a named entity recognition task, which is an important task for identifying meaningful terms and multiword phrases in free text [27]. In this study, we fine-tuned several BERT-based models for the purpose of identifying potential FH mentions from an EHR corpus, leveraging various data sets for the purpose of providing more context for fine-tuning BERT-based models. Lexicon entries were collected from a large clinical EHR corpus, mitigating the problem of missing entities caused by limited amounts of data. The dictionary coverage evaluation showed that it covers a greater range of lexical variants and focuses primarily on clinical concepts typically reported as part of FH relative to a direct lexicon generated from the UMLS and SNOMED-CT. Our corpus-driven lexicon features FH definitions based on semantic types, concept normalization to UMLS and SNOMED-CT CUIs, and manual curation, with the potential to resolve semantic ambiguity and promote interoperability among various systems. The rule-based FH system also provides standard Health Level 7 FHIR output to foster interoperability.

FH relation extraction is more relevant for downstream analysis compared with entity extraction. In previous challenges, the F1 score was regarded as the most important metric for relation extraction evaluation. The highest F1 score obtained from challenges was 0.5708 in the BioCreative challenge [19] and was 0.681 in the N2C2 challenge [20]. However, it is not our aim to compete with previous studies in terms of F1 scores. As relation extraction is still a challenging task, an FH system that prioritizes recall is highly desired for NLP-assisted curation in EHR-based studies. Our evaluation results showed that the rule-based FH system on top of the corpus-driven lexicon produced higher recall than that obtained with the deep learning–based FH system. In addition, the combined results from both the rule-based module and the deep learning–based FH system resulted in the highest recall for relation extraction, ranging from 0.7109 to 0.8370, which were higher than the recall values obtained in any previous challenge results, ranging from 0.3732 to 0.6810.

Note that we did not observe higher performance when using the reannotated data as compared to the original data. There may be two underlying reasons for this. First, reannotated data have not been used for training the rule-based system. Second, reannotated data were obtained by a professional annotator with deep domain knowledge, which makes the information extraction task harder. In addition, we observed that performance on 2019 N2C2 FH challenge data was worse than that on the 2018 BioCreative FH challenge data. This is mainly because the 2019 N2C2 FH challenge added a certainty attribute (negated or nonnegated) into the relation extraction, which made the relation extraction task harder.

Our FH synthetic data sets used for training and testing were from real clinical sentences, for which the observations, FMs, and ethnicities are shuffled among the whole corpus using a heuristic deidentification process. The granularity of the synthetic FH data sets is the same as that of real FH data. In this study, we have not exhaustedly compared all BERT-related models. Theoretically, large language models–empowered knowledge engineering is sufficient for lexicon entry collection from a clinical corpus. Our focus in this study was to provide a corpus-driven lexicon resource that leads to a rule-based FH baseline system for high-throughput analysis, while doing so in a manner that promotes interpretability and explainability for downstream applications.

We recommend that a comprehensive FH system include both a rule-based module and a deep learning–based module to obtain higher recall, which could facilitate manual curation. Although only the rule-based FH system can output normalized concepts, output from a deep learning–based FH system can be a rich source to enrich the lexicon. In the future, we will repeat the SFCP normalization for the output from the deep learning–based FH system to consistently improve the FH lexicon and the rule-based FH system. Meanwhile, we will continue to manually review the BERT-extracted entities without automatic

normalization with frequency under 20, and look into other data sources such as social media so as to expand more concepts and/or term variants for the current lexicon. In addition, we will engage the user community to continuously refine the lexicon. We also plan to update the lexicon and rule-based FH system yearly, which will be distributed through the same open-source repository on GitHub.

There are three limitations to this study. First, although a large corpus with 83,000 patients was used for collection of potential FH variants, there is still a possibility that the FH information is not well represented. In addition, as the lexicon was developed using a largely monoinstitutional data resource, the lexicon may not be generalizable in other institutions. Second, during the entity normalization, we simply adopted an existing standardization framework, as it was not a priority of this study to focus on standardization method development. Third, we arbitrarily set a frequency cutoff of 20 for entities that were not automatically normalized to include in our manual review. However, we realize that some entities with low frequency also have the potential to contribute to the lexicon entries, such as "rectal ca" with a frequency of 18 and "highcholesterol" with a frequency of 12.

In summary, we constructed a corpus-driven FH lexicon to serve as a language resource for FH information extraction. Standardization of concepts in the FH lexicon and the rule-based FH system foster interoperability. The resulting lexicon and the rule-based FH system are freely available as part of the OHNLP Toolkit ecosystem. In the future, we will continue to expand more concepts and/or term variants of the current lexicon, and will explore the incorporation of the system for data curation efforts needed in various EHR-based studies and applications.

Acknowledgments

This work was made possible by National Institutes of Health (NIH) grants (1U01TR002062-01, U24CA194215-01A1, and R01LM013240).

Data Availability

The source code underlying this article is available on GitHub [25].

Authors' Contributions

LW designed the study, implemented models, performed data analysis and manual review, and wrote the manuscript. SM and KJP conducted automatic normalization. HH implemented the models and performed data analysis. AW implemented the baseline system and revised the manuscript. SF revised the manuscript. XA, SL, and RK implemented the models. HL conceptualized and designed the study, performed data analysis, and critically revised the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Statistical comparison between original and enhanced annotations.
[DOCX File , 17 KB - medinform_v11i1e48072_app1.docx]

Multimedia Appendix 2
Degree of consanguinity.
[DOCX File , 14 KB - medinform_v11i1e48072_app2.docx]

Multimedia Appendix 3
Performance of various BERT models fine-tuned on different data sets.
[DOCX File , 15 KB - medinform_v11i1e48072_app3.docx]

References

1. Ginsburg GS, Wu RR, Orlando LA. Family health history: underused for actionable risk assessment. Lancet 2019 Aug 17;394(10198):596-603 [FREE Full text] [doi: 10.1016/S0140-6736(19)31275-9] [Medline: 31395442]
2. Bylstra Y, Lim WK, Kam S, Tham KW, Wu RR, Teo JX, et al. Family history assessment significantly enhances delivery of precision medicine in the genomics era. Genome Med 2021 Jan 07;13(1):3 [FREE Full text] [doi: 10.1186/s13073-020-00819-1] [Medline: 33413596]
3. Friedlin J, McDonald CJ. Using a natural language processing system to extract and code family history data from admission reports. AMIA Annu Symp Proc 2006;2006:925 [FREE Full text] [Medline: 17238544]
4. Polubriaginof F, Tatonetti NP, Vawdrey DK. An assessment of family history information captured in an electronic health record. AMIA Annu Symp Proc 2015;2015:2035-2042 [FREE Full text] [Medline: 26958303]

5. Fu S, Wang L, Moon S, Zong N, He H, Pejaver V, et al. Recommended practices and ethical considerations for natural language processing-assisted observational research: A scoping review. Clin Transl Sci 2023 Mar;16(3):398-411 [FREE Full text] [doi: 10.1111/cts.13463] [Medline: 36478394]
6. Wang L, Fu S, Wen A, Ruan X, He H, Liu S, et al. Assessment of electronic health record for cancer research and patient care through a scoping review of cancer natural language processing. JCO Clin Cancer Inform 2022 Aug;6:e2200006 [FREE Full text] [doi: 10.1200/CCI.22.00006] [Medline: 35917480]
7. Fu S, Chen D, He H, Liu S, Moon S, Peterson KJ, et al. Clinical concept extraction: a methodology review. J Biomed Inform 2020 Sep;109:103526 [FREE Full text] [doi: 10.1016/j.jbi.2020.103526] [Medline: 32768446]
8. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016 Mar 15;3:160018. [doi: 10.1038/sdata.2016.18] [Medline: 26978244]
9. Fu S. TRUST: Clinical text retrieval and use towards scientific rigor and transparent process. Thesis. University of Minnesota. 2021. URL: https://conservancy.umn.edu/handle/11299/226410 [accessed 2023-04-15]
10. Moon S, He H, Liu H. Sublanguage characteristics of clinical documents. 2022 Presented at: IEEE International Conference on Bioinformatics and Biomedicine (BIBM)2022; December 6-8, 2022; Las Vegas, NV, and Changsha, China. [doi: 10.1109/bibm55620.2022.9995620]
11. Johnson SB. A semantic lexicon for medical language processing. J Am Med Inform Assoc 1999;6(3):205-218 [FREE Full text] [doi: 10.1136/jamia.1999.0060205] [Medline: 10332654]
12. Luo Z, Duffy R, Johnson S, Weng C. Corpus-based approach to creating a semantic lexicon for clinical research eligibility criteria from UMLS. Summit Transl Bioinform 2010 Mar 01;2010:26-30 [FREE Full text] [Medline: 21347142]
13. Liu H, Wu ST, Li D, Jonnalagadda S, Sohn S, Wagholikar K, et al. Towards a semantic lexicon for clinical natural language processing. AMIA Annu Symp Proc 2012;2012:568-576 [FREE Full text] [Medline: 23304329]
14. Devlin J, Chang M, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv. 2018. URL: https://arxiv.org/abs/1810.04805 [accessed 2023-04-15]
15. Clark K, Khandelwal U, Levy O, Manning C. What does BERT look at? an analysis of BERT's attention. arXiv. 2019. URL: https://arxiv.org/abs/1906.04341 [accessed 2023-04-15]
16. Min S, Seo M, Hajishirzi H. Question answering through transfer learning from large fine-grained supervision data. arXiv. 2017. URL: https://arxiv.org/abs/1702.02171 [accessed 2023-04-15]
17. Bodenreider O, McCray AT. Exploring semantic groups through visual approaches. J Biomed Inform 2003 Dec;36(6):414-432 [FREE Full text] [doi: 10.1016/j.jbi.2003.11.002] [Medline: 14759816]
18. Liu H, Bielinski SJ, Sohn S, Murphy S, Wagholikar KB, Jonnalagadda SR, et al. An information extraction framework for cohort identification using electronic health records. AMIA Jt Summits Transl Sci Proc 2013;2013:149-153 [FREE Full text] [Medline: 24303255]
19. Liu S, Mojarad M, Wang Y. Overview of the BioCreative/OHNLP family history extraction task. ResearchGate. 2018. URL: https://www.researchgate.net/publication/327424806_Overview_of_the_BioCreativeOHNLP_2018_Family_History_Extraction_Task#fullTextFileContent [accessed 2023-06-07]
20. Shen F, Liu S, Fu S, Wang Y, Henry S, Uzuner O, et al. Family history extraction from synthetic clinical narratives using natural language processing: overview and evaluation of a challenge data set and solutions for the 2019 National NLP Clinical Challenges (n2c2)/Open Health Natural Language Processing (OHNLP) Competition. JMIR Med Inform 2021 Jan 27;9(1):e24008 [FREE Full text] [doi: 10.2196/24008] [Medline: 33502329]
21. Uzuner, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. J Am Med Inform Assoc 2011;18(5):552-556 [FREE Full text] [doi: 10.1136/amiajnl-2011-000203] [Medline: 21685143]
22. Alsentzer E, Murphy J, Boag W. Publicly available clinical BERT embeddings. arXiv. 2019. URL: https://arxiv.org/abs/1904.03323 [accessed 2023-04-15]
23. Michalopoulos G, Wang Y, Kaka H, Chen H, Wong A. Umlsbert: Clinical domain knowledge augmentation of contextual embeddings using the unified medical language system metathesaurus. arXiv. 2020. URL: https://arxiv.org/abs/2010.10391 [accessed 2023-04-15]
24. Peterson KJ, Jiang G, Liu H. A corpus-driven standardization framework for encoding clinical problems with HL7 FHIR. J Biomed Inform 2020 Oct;110:103541 [FREE Full text] [doi: 10.1016/j.jbi.2020.103541] [Medline: 32814201]
25. OHNLP. Rule-based family history NLP system built on lexicon. arXiv. 2023. URL: https://github.com/OHNLP/FamilyHistoryNLP [accessed 2023-04-15]
26. Zhong Z, Chen D. A frustratingly easy approach for entity and relation extraction. arXiv. 2020. URL: https://arxiv.org/abs/2010.12812 [accessed 2023-04-15]
27. Campos D, Matos S, Oliveira J. Biomedical named entity recognition: a survey of machine-learning tools. In: Sakurai S, editor. Theory and applications for advanced text mining. London, UK: IntechOpen; 2012.

Abbreviations

BERT: bidirectional encoder representations from transformers

CUI: concept unique identifier
EHR: electronic health record
FAIR: Findable, Accessible, Interoperable, and Reusable
FH: family history
FHIR: Fast Healthcare Interoperability Resource
FM: family member
FN: false negative
N2C2: 2019 NLP Clinical Challenge
NLP: natural language processing
OHNLP: Open Health Natural Language Processing
PURE: Princeton University Relation Extraction
RITE: Reproducible, Implementable, Transparent, and Explainable
SFCP: standardization framework for clinical problems
SNOMED-CT: Systematized Nomenclature of Medicine Clinical Terms
TP: true positive
UMLS: Unified Medical Language System

Edited by G Eysenbach, C Lovis; submitted 13.04.23; peer-reviewed by X Wang, J Xia, J Zheng, M Torii; comments to author 12.05.23; revised version received 25.05.23; accepted 01.06.23; published 27.06.23.

Please cite as:
Wang L, He H, Wen A, Moon S, Fu S, Peterson KJ, Ai X, Liu S, Kavuluru R, Liu H
Acquisition of a Lexicon for Family History Information: Bidirectional Encoder Representations From Transformers–Assisted Sublanguage Analysis
JMIR Med Inform 2023;11:e48072
URL: https://medinform.jmir.org/2023/1/e48072
doi:10.2196/48072
PMID:37368483

Original Paper

A Multilabel Text Classifier of Cancer Literature at the Publication Level: Methods Study of Medical Text Classification

Ying Zhang[1*], MA; Xiaoying Li[1*], PhD; Yi Liu[1], MA; Aihua Li[1], PhD; Xuemei Yang[1], PhD; Xiaoli Tang[1], MA

Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing, China

*these authors contributed equally

Corresponding Author:
Xiaoli Tang, MA
Institute of Medical Information
Chinese Academy of Medical Sciences
No 69, Dongdan North Street
Beijing, 100020
China
Phone: 86 10 52328902
Email: tang.xiaoli@imicams.ac.cn

Related Article:

This is a corrected version. See correction statement: http://medinform.jmir.org/2024/1/e62757/

Abstract

Background: Given the threat posed by cancer to human health, there is a rapid growth in the volume of data in the cancer field and interdisciplinary and collaborative research is becoming increasingly important for fine-grained classification. The low-resolution classifier of reported studies at the journal level fails to satisfy advanced searching demands, and a single label does not adequately characterize the literature originated from interdisciplinary research results. There is thus a need to establish a multilabel classifier with higher resolution to support literature retrieval for cancer research and reduce the burden of screening papers for clinical relevance.

Objective: The primary objective of this research was to address the low-resolution issue of cancer literature classification due to the ambiguity of the existing journal-level classifier in order to support gaining high-relevance evidence for clinical consideration and all-sided results for literature retrieval.

Methods: We trained a multilabel classifier with scalability for classifying the literature on cancer research directly at the publication level to assign proper content-derived labels based on the "Bidirectional Encoder Representation from Transformers (BERT) + X" model and obtain the best option for X. First, a corpus of 70,599 cancer publications retrieved from the Dimensions database was divided into a training and a testing set in a ratio of 7:3. Second, using the classification terminology of International Cancer Research Partnership cancer types, we compared the performance of classifiers developed using BERT and 5 classical deep learning models, such as the text recurrent neural network (TextRNN) and FastText, followed by metrics analysis.

Results: After comparing various combined deep learning models, we obtained a classifier based on the optimal combination "BERT + TextRNN," with a precision of 93.09%, a recall of 87.75%, and an F_1-score of 90.34%. Moreover, we quantified the distinctive characteristics in the text structure and multilabel distribution in order to generalize the model to other fields with similar characteristics.

Conclusions: The "BERT + TextRNN" model was trained for high-resolution classification of cancer literature at the publication level to support accurate retrieval and academic statistics. The model automatically assigns 1 or more labels to each cancer paper, as required. Quantitative comparison verified that the "BERT + TextRNN" model is the best fit for multilabel classification of cancer literature compared to other models. More data from diverse fields will be collected to testify the scalability and extensibility of the proposed model in the future.

(JMIR Med Inform 2023;11:e44892) doi:10.2196/44892

KEYWORDS

text classification; publication-level classifier; cancer literature; deep learning

Introduction

Background

According to the World Health Organization (WHO) reports, cancer is one of the leading causes of death worldwide, accounting for nearly 10 million deaths in 2020 [1]. With cancer emerging as the greatest threat to human life, there is a rapid growth in the volume of literature published in the cancer field. The trend of disciplinary convergence has led to publications requiring labels from multiple subjects. Consequently, there is increasingly more demand for accurate cancer literature classification for retrieval, evidence support, academic analysis, and statistical evaluation in order to support clinical research, precision medicine, and discovery of interdisciplinary cancer research [2,3] by forecasting trends and hotspot statistics.

Text classification is the process of assigning specific labels to the literature based on individual features. The current methods of classifying the literature can be divided into 3 groups: mapping based, subject information based, and machine learning based [4-6]. Recently, an increasing number of studies have experimented with deep learning to enhance the effects of text classification [4]. Most of the existing literature classification (eg, Web of Science, Scopus [7-11]) is usually carried out at the journal level, that is, all papers in a given journal get the same labeling categories as the hosted journal. However, given that interdisciplinary research is increasing in the cancer field [12], there is a need for a more precise classifier, as papers from a journal always present a diverse range of topics [13,14]. Moreover, literature classification at the journal level can no longer adapt to the dynamics of newly developing subjects and fully characterize text features.

Related Works

Development of Text Classification Technology

Text classification technology has undergone rapid development from expert systems to machine learning to finally deep learning [15]. Maron [16] published the first paper on automatic text classification in 1961. By the end of the 20th century, machine learning had matured into a fully developed field [17]. Joachims [18] established the bag-of-words model to transform text into a vector with a fixed length and then selected features using the information gain criterion to achieve dimensionality reduction, eventually training the feature vector iteratively using a support vector machine (SVM) classifier. Rasjid [19] focused on data classification using k-nearest neighbors (k-NN) and naive Bayes. Liang et al [20] improved the feature recognition of the literature using appropriate clusters and the introduction of differential latent semantics index (DLSI) spaces. In 2006, with the rapid development of deep learning, text classification research based on deep learning gradually replaced traditional machine learning methods and became the mainstream, with wide applications in numerous tasks [4].

The deep learning–based text classification method adopts word vectors (eg, GloVe [21] and word2vec [22]) for word semantic representation [23], and subsequently, various deep neural network (DNN)–based text classification methods develop. Convolutional neural networks (CNNs) [24] were originally constructed for image processing and have been broadly used for text classification [25-27]. Due to the computational complexity, deep network gradient disappearance, and short text content of a CNN, a series of optimized models were gradually derived to address these issues, including FastText [28], deep pyramid convolutional neural networks (DPCNNs) [29], knowledge pyramid convolutional neural networks (KPCNNs) [30], and text convolutional neural networks (TextCNNs) [26]. Especially, the TextCNN is a simple, shallow network and requires only a small number of hyperparameters for fine-tuning. Compared with CNNs, recurrent neural networks (RNNs) can easily implement multilayer superposition to construct a multilayer neural network [31], such as multilayer long short-term memory (LSTM) [32] or multilayer gate recurrent unit (GRU). In terms of improvements to RNNs, the text recurrent neural network (TextRNN) uses a multitask learning framework to jointly learn across multiple related tasks; a deep recurrent neural network (DRNN) [23] incorporates position invariance into an RNN, captures local features to find the optimal window size, and then achieves marked improvements over RNN and CNN models. All these models have laid the foundation for follow-up studies.

Bidirectional Encoder Representation from Transformers (BERT) has emerged as a new linguistic representation model by the introduction of attentional mechanisms, which have been broadly applied in machine translation [32], image description generation [33], machine reading comprehension [34], and text classification [35]. Being a bidirectional encoder model based on a transformer, BERT became an important advancement of natural language processing, especially text classification. For example, Shen et al [36] attempted to train a Chinese corpus–based $BERT_{base}$ model for the classification of the literature on Chinese social science and technology and also explored its application to practical production. In addition, Lu and Ni [37] developed a multilayer model for patent classification using the combination "BERT + CNN," while Liu et al [38] proposed a sentence-BERT for hierarchical clustering of literature abstracts. However, the accuracy of applying universal language models directly to the biomedical field is not sufficient, and this motivated studies to train the biomedical BERT from scratch. Typical instances are BioBERT pretrained on PubMed citations and PubMed Central (PMC) full text [39] and PubMedBERT obtained with mixed-domain pretraining using PubMed text and clinical notes [40]. Up to now, BioBERT and PubMedBERT have achieved success in named entity recognition, extraction of relationships between entities, entity normalization [41,42], *International Classification of Diseases* (ICD) autocoding [43], and its multilabel classification (MLC) [44]. These studies afford us lessons that merit attention.

Multilabel Text Classification

The deep learning model has contributed to the success of MLC due to its dynamic representation learning and end-to-end learning framework. Multilabel text classification (MLTC) is the application of MLC to the task of text classification and the assigning of a set of targeted labels to each sample [45], which has been part of the longstanding challenge in both academia and industry. In the biomedical domain, Du et al [46] proposed an end-to-end deep learning model ML-Net for biomedical text, while Glinka et al [47] focused on a mixture of feature selection methods, filter, and wrapper methods. In addition, Hughes et al [48] tried to classify medical text fragments at the sentence level based on a CNN, and Yogarajan et al [49] used a multilabel variant of medical text classification to enhance the prediction of concurrent medical codes. Automatic question-and-answer systems and auxiliary decision-making are the 2 typical applications of MLTC. For example, Wasim et al [50] proposed a classification model for multilabel problems in the flow of biomedical question-and-answer systems with factlike and listlike questions. Similarly, Baumel et al [51] presented a hierarchical attentional bidirectional gated cyclic unit that used attention weights to better understand the sentence and word with the greatest impact on the decision.

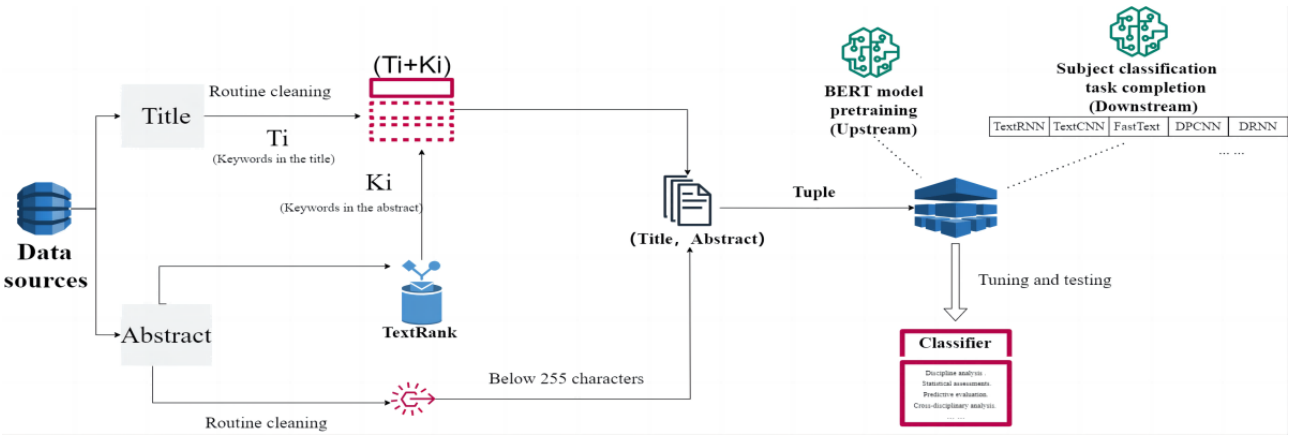
In this study, we adopted MLTC technology to classify cancer publications for better retrieval, academic analysis, and statistical evaluation. We introduced a method of multilabel classification for cancer research at the publication level based on the "BERT + X" model. BERT is a learning model that can migrate to other tasks to obtain better outcomes, which was pretrained on a large and easily accessible data set, and X is a deep learning model that can capture semantic features of text accurately. The combined model was trained on a corpus from a multilabel publication database, and then cancer publications were classified into appropriate categories directly at the publication level instead of the journal level.

Methods

Study Design

This study mainly aimed to train a deep learning–based multilabel classifier for cancer literature classification at the publication level. The overall framework of the study is illustrated in Figure 1. First, a corpus of titles and abstracts of cancer publications retrieved from the Dimensions database were divided into a training and a testing set in a ratio of 7:3 after preprocessing. Second, to capture sufficient text features for multilabel classification, the titles and abstracts were taken separately as 2 independent layers, which were called the "tuple" in this study. Finally, "BERT + X" classifiers based on 5 deep learning models were trained; X refers to "TextRNN," "TextCNN," "FastText," "DPCNN," and "DRNN." The performance of the candidate classifiers was compared quantitatively in terms of 3 conventional metrics in order to identify the optimal model for classifying cancer literature.

Figure 1. Study framework. BERT: Bidirectional Encoder Representation from Transformers; DPCNN: deep pyramid convolutional neural network; DRNN: deep recurrent neural network; TextRNN: text recurrent neural network; TextCNN: text convolutional neural network.



Data Collection and Preprocessing

Refined data with semantic and context features are the basis of deep learning model training. In this study, to train a lightweight, compatible, and high-applicability classifier for text classification, we first preprocessed cancer literature data to extract features and sequential semantic information. The classification terminology maintained by the International Cancer Research Partnership (ICRP), called cancer type (CT), was used as standard labels to characterize individual cancer studies in terms of 62 CTs. The ICRP CT has been linked to the ICD maintained by WHO [52] and is increasingly gaining rapid recognition worldwide. Moreover, the ICRP CT has been applied to several international databases for labeling cancer literature and research documents with fine granularity. A typical example is the Dimensions database (Digital Science & Research Solutions, Inc), which covers more than 135 million publications, 6.5 million grants, and 153 million patents, providing a collaborative path to enhanced scientific discovery with transparent data sources. Importantly, cancer publications with ICRP CT–classified labels from the Dimensions database provide a way to prepare annotation data for model training.

Construction of the Corpus and Balanced Sampling

A set of 70,599 publications from 2003 to 2022 was randomly sampled from the Dimensions database using the keyword "cancer," along with the ICPR CT labels for each publication. Figure 2 shows the distribution of different CTs among the corpus data. Here, to intuitively demonstrate the volume distribution in the 62 different cancer categories, we ranked the

categories in descending order of the number of corresponding publications included. Categories with sample sizes larger than 500 were listed separately, while the remaining 41 (66.13%) categories were grouped into 3 classes: CT22-CT35, n=14 (34.15%) categories; CT36-CT49, n=14 (34.15%); and CT50-CT62, n=13 (31.71%). The top 9 (14.52%) categories (breast cancer, non-site-specific cancer, colon and rectal cancer, lung cancer, prostate cancer, ovarian cancer, stomach cancer, cervical cancer, and pancreatic cancer) accounted for 76.35% (53,899/70,599) of the total corpus. Obviously, the distribution of labeling data was uneven, and the top 9 labels contained more than three-fourths of the total corpus. Once the original corpus was directly used to train the deep learning model without balanced sampling, the resulting classifier would cause overfitting due to a few categories with excessive volume and fail to be generalized in practical use.

To avoid the degradation of precision caused by overfitting, we balanced the data sampling before model training. First, we ranked the cancer categories in descending order of the number of corresponding publications. Next, a threshold (70 in our study) used for setting the sampling index was obtained after multiple testing, followed by calculation of the index using categories with more than 70 publications. The resulting mean , median, variance, standard deviation of the number of samples contained in each category were recorded as follows: mean=1127, median=282, variance=5,889,927, and standard deviation=2427. Here, the median was selected as the initial index according to the actual distribution of sampling data. Considering that 30% of the corpus was used as the testing set, the final index was set to 500 in terms of the number of publications for a competent training set. Therefore, categories whose number of publications exceeded the final index were balanced and down-sampled separately, which means that 500 publications from each category were randomly extracted for a uniform corpus.

Table 1 shows part of the balanced sampling results. It is clear that the optimized corpus contributed to a reduction in the adverse effects of overfitting and was subsequently used for keyword extraction and model training.

Figure 2. Original sample distribution of CTs. CT: cancer type.

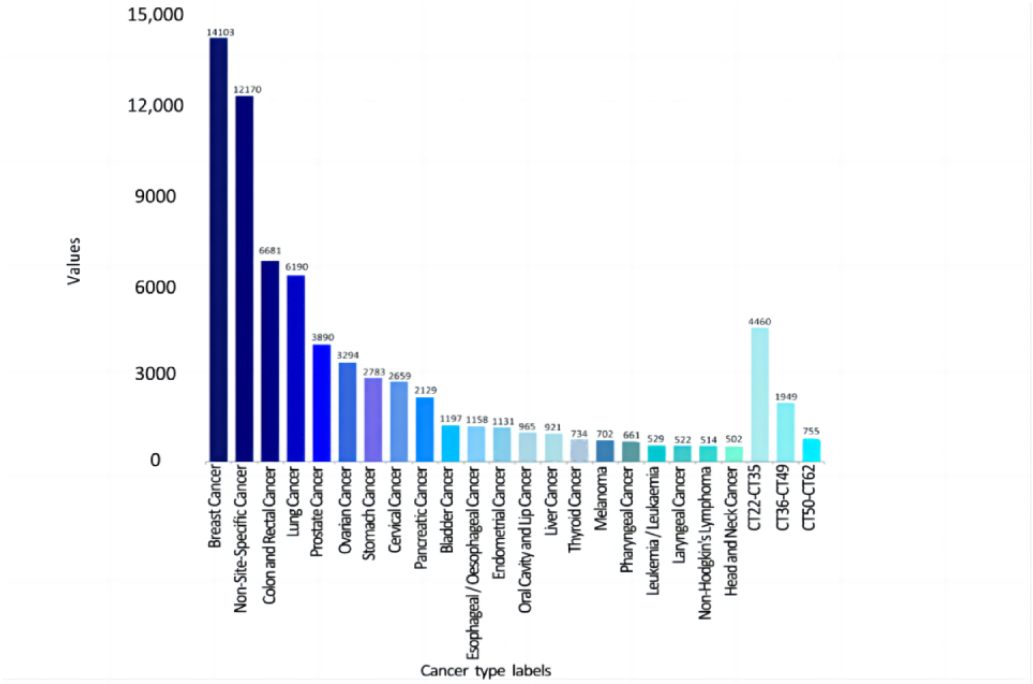


Table 1. Example of balanced sampling of top 21 categories (N=70,599 publications).

Sequence number	ICRP CT[a]	ICRP code[b]	ICD-10[c] code	Original sample[d], n (%)	Balanced sample[e], n (%)
1	Breast cancer	7	C50	14,103 (19.98)	500 (3.55)
2	Non-site-specific cancer	2	N/A[f]	12,170 (17.24)	500 (4.11)
3	Colon and rectal cancer	64	C18, C19, C20	6681 (9.46)	500 (7.48)
4	Lung cancer	28	C34, C45	6190 (8.77)	500 (8.08)
5	Prostate cancer	42	C61	3890 (5.51)	500 (12.85)
6	Ovarian cancer	66	C56	3294 (4.67)	500 (15.18)
7	Stomach cancer	51	C16	2783 (3.94)	500 (17.97)
8	Cervical cancer	9	C53	2659 (3.77)	500 (18.80)
9	Pancreatic cancer	37	C25	2129 (3.02)	500 (23.49)
10	Bladder cancer	3	C67	1197 (1.70)	500 (41.77)
11	Esophageal/oesophageal cancer	12	C15	1158 (1.64)	500 (43.18)
12	Endometrial cancer	11	C54	1131 (1.60)	500 (44.21)
13	Oral cavity and lip cancer	36	C00, C01, C02, C03, C04, C05, C06, C09	965 (1.37)	500 (51.81)
14	Liver cancer	23	C22	921 (1.30)	500 (54.29)
15	Thyroid cancer	54	C73	734 (1.04)	500 (68.12)
16	Melanoma	29	C43	702 (0.99)	500 (71.23)
17	Pharyngeal cancer	61	C14.0	661 (0.94)	500 (75.64)
18	Leukemia/leukaemia	27	C91, C92, C93, C94, C95	529 (0.75)	500 (94.52)
19	Laryngeal cancer	26	C32	522 (0.74)	500 (95.79)
20	Non-Hodgkin's lymphoma	35	C82, C83, C84, C85, C96.3	514 (0.73)	500 (97.28)
21	Head and neck cancer	21	C76.0	502 (0.71)	500 (99.60)

[a]ICRP CT: International Cancer Research Partnership Cancer Type; here, "ICRP CT" denotes the label.

[b]"ICRP code" refers to the label code.

[c]ICD-10: International Classification of Diseases, Tenth Revision; this is the ICD code linked to the appropriate ICPR CT.

[d]"Original sample" represents the number of publications obtained directly from the Dimensions database.

[e]"Balanced sample" means the number of publications after balanced sampling.

[f]N/A: not applicable.

Construction of a Tuple Consisting of a Title and an Abstract

The corpus of cancer publications consisted of titles and abstracts in English, while the title and abstract of each publication was saved separately for ease of use. Generally, the title is a short sentence of a confined length to express an independent meaning, which has a high rate of conformity regarding the content. Comparatively, an abstract is also valuable, since it clearly and accurately summarizes the main content of the publication by expressing its purpose, methods, results, and conclusions. In this study, both title and abstract were independently used to train a 2-layer classifier based on their semantic and context features, called a tuple for simplicity.

Keyword Extraction From the Abstracts of Cancer Publications

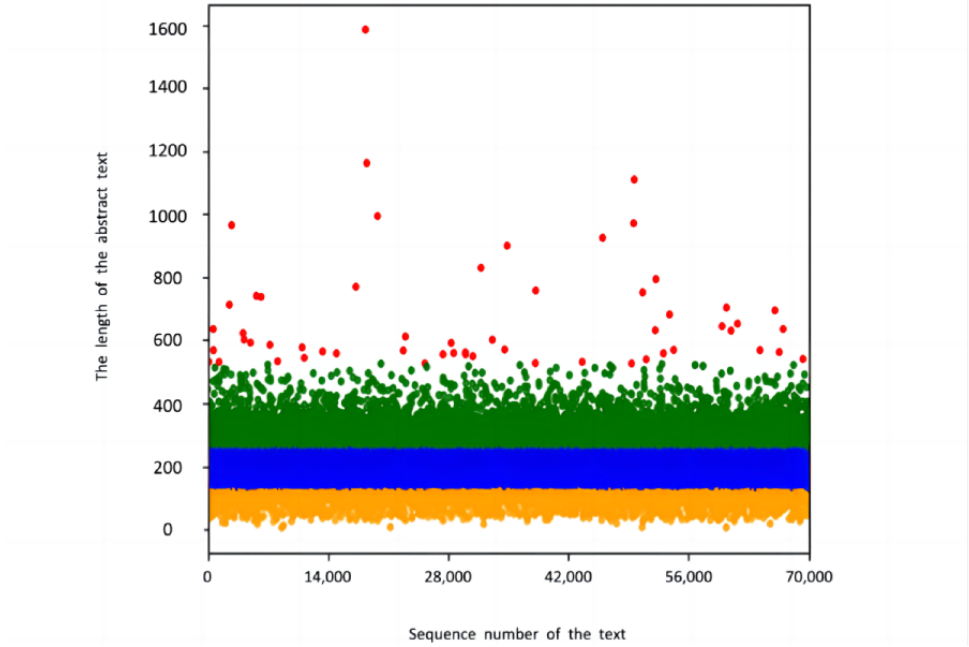
In this study, in contrast to the abstract layer, the operation of the title layer mainly focused on keyword training. However, the number of valid keywords contained in titles is quite limited. To improve feature representation and model performance, more keywords were extracted from the abstracts and merged with the title layer. The TextRank algorithm [53] was adopted for keyword extraction from the abstracts, taking advantage of the co-occurring semantics information between words from the given sentences.

A lightweight classifier was desired in this study, so the length of the abstracts needed to be controlled to keep a balance between the running speed, effectiveness, and volume occupied. To ascertain the most appropriate abstract length, we analyzed the scatter plot of the abstract length distribution, as shown in Figure 3. Here, the horizontal axis represents the serial number of publications, the vertical axis represents the length of the publication abstracts, and the red (length>512), green (256<length≤512), blue (128<length≤256), and orange (length≤128) colors represent different abstract lengths separately. Among them, the blue zone was evenly distributed

and had a high proportion, the red zone had the least proportion, and the green and orange zones were comparable.

After statistical analysis, the maximum length of the abstracts was set to 256 characters mainly for 3 reasons. First, 256 is the 8th power of 2, which facilitates machine understanding after tuning [53]. Second, we wanted to collect the most valid information as much as possible, while reducing sparsity. Third, we wanted to avoid the learning of shallow networks, which are equivalent to follow-up layers when training (ie, gradient disappearance).

Figure 3. Scatter plot of distribution of abstract length.



Training a Model for MLCT

Upstream Pretrained Language Models

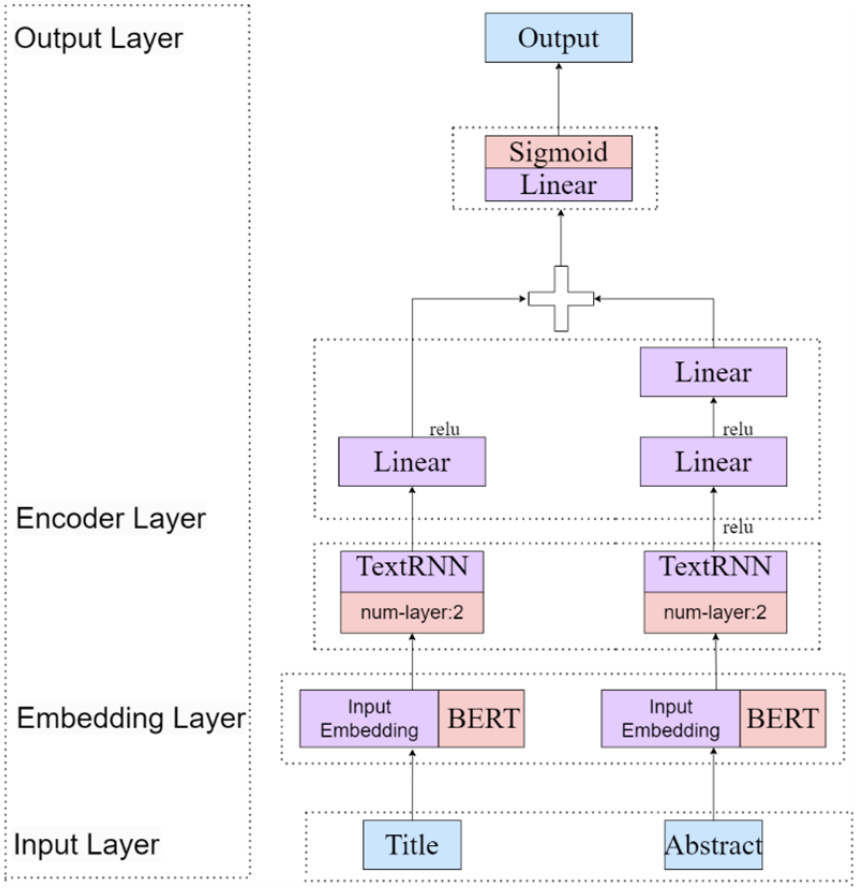
Pretraining generally refers to putting a large amount of low-cost collected training data together, learning the commonalities of the data, and then tuning the model that has the commonalities with a small amount of labeled data of a specific domain. Therefore, pretrained language models start from the commonalities and learn the special parts of the specific task. BERT is a successful model pretrained on Wikipedia and a book corpus via self-supervision tasks, and fine-tuning benefits downstream tasks. Being a pretrained language model based on the bidirectional transformer encoder architecture, BERT uses sentence-level negative sampling to obtain sentence representation/sentence pair relationships. In addition, BERT takes advantage of the transformer model instead of LSTM for expressive and temporal efficiency, as well as the masked language model to extract contextual features. We used BERT to handle specific natural language processing tasks downstream to produce word vectors in the pretraining stage. During the fine-tuning, we then completed data training for the pretrained BERT through the output layer based on cancer publications in order to save time and improve accuracy.

Downstream Classification Model Training

To complete the actual task of downstream natural language processing based on the fine-tuned upstream procedure, we tried to train several preliminary language models and then chose one as the optimal model for our classifier according to its all-round performance. The “BERT + X” pattern was adopted for the classifier to determine the optimum option for X and the best way to combine BERT and X. According to the actual scenario and expert consultations, including the length, tightness of context, and multidisciplinarity, 5 models suitable for cancer publications were compared for the classification model: TextCNN, TextRNN, FastText, DPCNN, and DRNN. Eventually, the definitive combined classifier would come out dependent on the comprehensive performance analysis of these 5 models.

The structure of the classification model is shown in Figure 4, where the TextRNN is selected as a representation of the 5 models, for example. Here, the title and abstract were input into the title layer and the abstract layer, respectively, while the word vectors were converted by BERT and passed to the encoder layer. The word vectors output from the title and abstract layers were stitched together and then transferred to the fully connected layer for normalization, while the final output of multilabel classification was generated by the sigmoid layer with the activation function.

Figure 4. Structure of the classification model. BERT: Bidirectional Encoder Representation from Transformers; TextRNN: text recurrent neural network.



Testing and Verification

To evaluate the performance of the trained classifier, a subset of the testing sample was selected from the corpus, which covered all 62 classification labels defined in the ICRP CT. We applied 3 frequently used indexes, namely precision, recall, and the F_1-score, to verify the classification results of the 5 models. Here, the F_1-score is the harmonic mean of precision and recall; the larger the F_1-score, the better the performance of the classification model. The quantitative indexes of the 5 "BERT + X" models were compared numerically and independently to choose X.

Results

Quantitative Analysis of the Performance of Classification Models

The test results of the combined classification downstream models are shown in Table 2, where we compared performance on 3 aspects (precision, recall, F_1-score) of 5 classification models (BERT + TextRNN, BERT + TextCNN, BERT + FastText, BERT + DPCNN, and BERT + DRNN). All metrics of BERT + TextRNN were consistently at a high level, with a precision of 93.09%, a recall of 87.75%, and an F_1-score of 90.34%. Here, BERT was directly used for fine-tuning training, combined with the TextRNN for multilabel classification. After adjusting and testing the parameters several times, the best parameters were obtained and are shown in Table 3.

Table 2. Performance comparison of 5 different "BERT[a] + X" models.

Model	Precision (%)	Recall (%)	F_1-score (%)
BERT + TextRNN[b]	93.09	87.75	90.34
BERT + TextCNN[c]	84.19	79.69	81.88
BERT + FastText	93.05	75.73	83.50
BERT + DPCNN[d]	81.78	75.00	78.25
BERT + DRNN[e]	88.98	53.05	66.47

[a]BERT: Bidirectional Encoder Representation from Transformers.

[b]TextRNN: text recurrent neural network.

[c]TextCNN: text convolutional neural network.

[d]DPCNN: deep pyramid convolutional neural network.

[e]DRNN: deep recurrent neural network.

Table 3. Parameters of the optimal "BERT[a] + TextRNN[b]" model.

Parameter	Value
Num_ train_ epochs	200.0
Max_ seq_ length	256
leaning_ rate	0.0001
train_ batch_ size	32
Predict_ batch_ size	32
Drop	0.5
Dense1	256
Dense2	62
TextRNN	256×2
LTSM[c]_UNITS	5
BERT_OUTDIM	768

[a]BERT: Bidirectional Encoder Representation from Transformers.

[b]TextRNN: text recurrent neural network.

[c]LSTM: long short-term memory.

Supplementary Analysis of the Model Structure

The proposed classification model takes the title and abstract of a publication as independent input, namely a tuple, as mentioned in the "Methods" section. To verify the effectiveness of the tuple input of the trained model, we conducted a set of comparison experiments based on "BERT + TextRNN." Especially, "2 tuples and 2 levels" represents the model that took the title and abstract as 2 levels of the training model separately, "1 unit and 1 level" represents the model that combined the title and abstract as whole-text input for training, and "the title alone" and "the abstract alone" represent the models that took the title or the abstract alone as input, respectively. Table 4 records the performance of different models from supplementary experiments, where the "2 tuples and 2 levels" model was superior, with a precision of 93.09%, a recall of 87.75%, and an F_1-score of 90.34%. The reason is that when applying the title or abstract alone to train the classification model, feature reduction occurs, which further leads to inferior performance of classification. In addition, compared with the "1 unit and 1 level" model taking the title and abstract as a 1-part text input, the "tuple and 2 levels" model enhanced the specificity of feature extraction. Notice that the title and abstract make different contributions to the subject of publication, and the classification model will lose the sufficient feature of the abstract if it is not trained separately. Eventually, the classification of the tuple input was elected for the proposed classification model. This is also identical to most of the subject-based literature processing, which take the title and abstract as independent text, such as subject indexing.

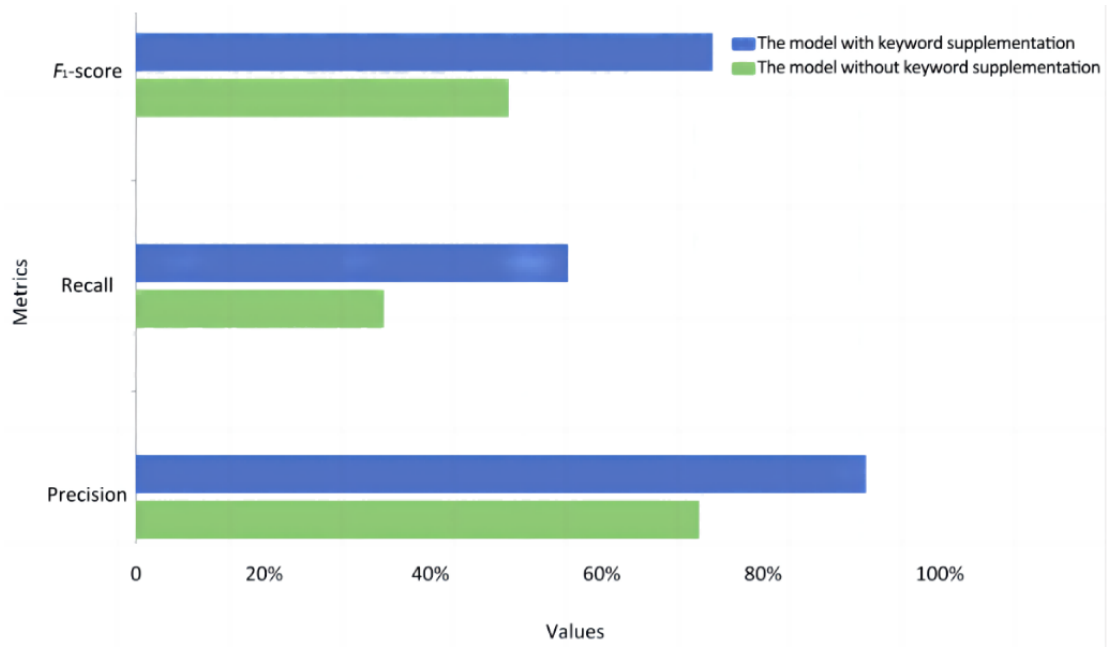
In addition, the proposed classification model used the TextRank algorithm to extract keywords from the abstract and supplement the title layer with them. To demonstrate the necessity and effectiveness of this step, Figure 5 shows the numerical indexes of the classification model with and without keyword extraction and supplementation. Here, the performance of the model with TextRank keyword extraction is shown in blue, and the results with direct training using titles and abstracts separately are plotted in green. It is obvious that the model without keyword

supplementation had a lower performance in recall, precision, and the F_1-score, which confirms the efficiency and effectiveness of the proposed classification model from a different perspective.

Table 4. Comparison of different types of input.

Input	Precision (%)	Recall (%)	F_1-score (%)
2 tuples and 2 levels	93.09	87.75	90.34
1 unit and 1 level	82.32	79.78	81.04
The title alone	44.37	31.54	36.88
The abstract alone	85.50	79.79	82.55

Figure 5. Comparison of the model with and without keyword supplementation.



Comprehensive Analysis of the Multilabel Classification

To explore whether there was a particular regularity in the distribution of multiple labels among different categories, the proportion of multilabel publications was statistically analyzed for classifier training. In total, 15,296 (21.67%) of the 70,599 publications had 2 or more labels, that is, more than one-fifth of the publications were multilabel ones. In addition, the categories were counted based on the characteristics of the number of labels (Figure 6) to visualize multilabel distributions. Here, we listed the top 20 categories by the volume of the publications included. The blue color refers to the total number of samples collected under a specific category, the green color is the number of samples with multiple labels under that category, and the yellow color denotes those with at least 3 labels. The categories with fewer samples had a higher ratio of multiple labels, and multiple labels had different characteristics among different categories. Analysis for a deeper relationship between multiple labels is necessary.

There are principally 2 roles of comprehensive analysis of multilabel publications. First, it highlighted the process of balanced sampling. Since part of the publications belonged to multiple categories and some of the categories had a high co-occurrence frequency compared to other categories, direct model training on the original corpus would lead to overfitting due to uneven distribution of samples. This is why we selected multilabel papers instead of those with a single label in order to obtain a balanced sample for classifier training. Second, the multilabel publications revealed the potential semantic correlation of texts, which provided a direction for the analysis of data characteristics. Based on the co-occurrence correlation and distribution between different categories, the semantic features were further characterized and the proposed classification model extended to other data with the same characteristics.

To explore the inherent correlation between multiple labels, we selected 2500 multilabel publications from the corpus for characteristics analysis. Specifically, samples with 2 labels accounted for 59.04% (1476/2500) and samples with at least 3 labels accounted for 40.96% (1024/2500) of all publications. Table 5 lists part of the analysis results. For instance, different CTs often co-occurred for statistical surveys in the literature review with a weak association.

The correlation strength of multiple labels of cancer publications were independently reviewed and assessed by 3 biocurators with relevant knowledge. Concretely, a publication with 2 labels and a clear semantic correlation within the corresponding subject classification labels was interpreted as “1,” while a publication

with 3 or more labels and more than two-third of the labels holding an obvious semantic association were also considered as "1". Once the 2 biocurators reached the same results, that specific publication was passed into the "review completed" data set. When they had different opinions, the corresponding publication was annotated as "pending review." After the first round of reviewing, the "pending review" data set was discussed together for the second time, and a third biocurator was invited for confirmation and agreement.

Figure 7 shows specific numbers of labeled publications with interlabel correlation, where the "strong association" zone consists of publications whose co-occurrence labels had explicit links between semantics, the "low association" zone consists of publications whose co-occurrence labels did not have clearly semantic links, and the "independent examples" zone consists of publications whose cancer labels were taken as single entities or independent examples for observation without intrinsic correlations. Of the 1476 publications with 2 labels, 1201 (81.37%) had a strong association, while 572/1024 (55.86%) publications with at least 3 labels had a low association. Among the publications with lowly correlated labels, 718/1024 (70.12%) took the different categories of cancers as a single entity or independent examples for observation without intrinsic correlations. We noticed some possible influence on the association distributions for training. In addition, the relationship between 2 labels in a publication was stronger than that among 3 or more labels, which justifies the demand to classify publications by subject at the publication level. Therefore, the strength of interlabel association could achieve the effect of assisting decision-making after multilabel classification to further support clinical diagnosis and treatment. In the future, we will carry out knowledge mining based on the existing interlabel semantic network and strengthen the training of interlabel association to improve the performance of the proposed classification model.

Figure 6. Comparison of samples with multilabels.

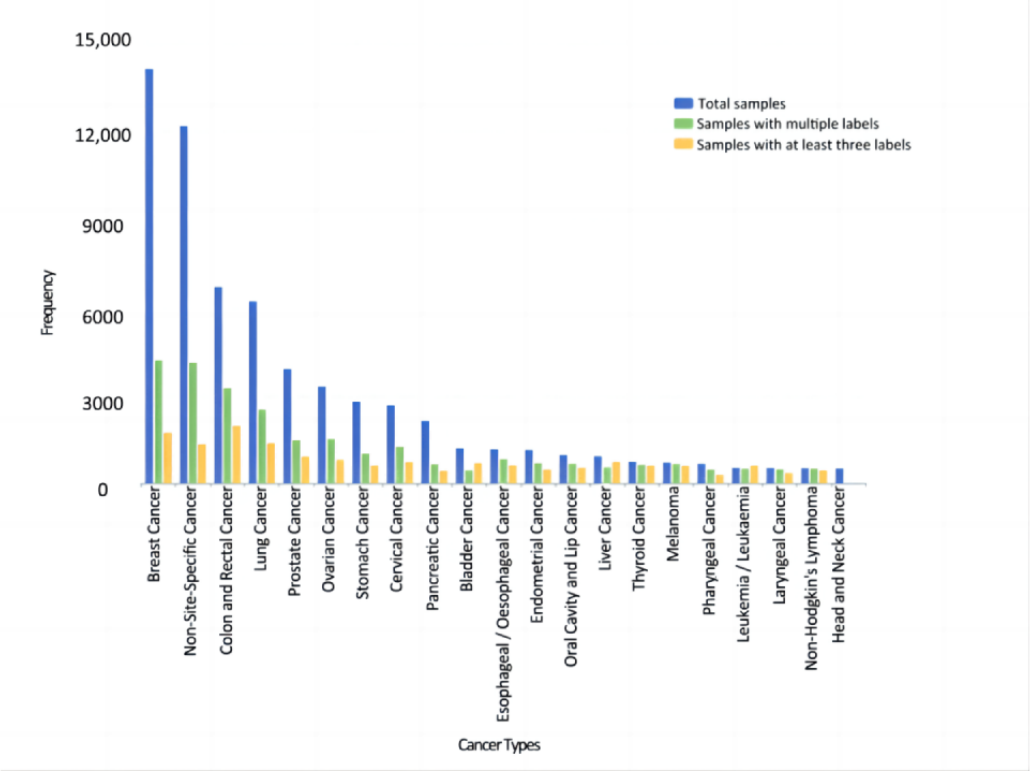
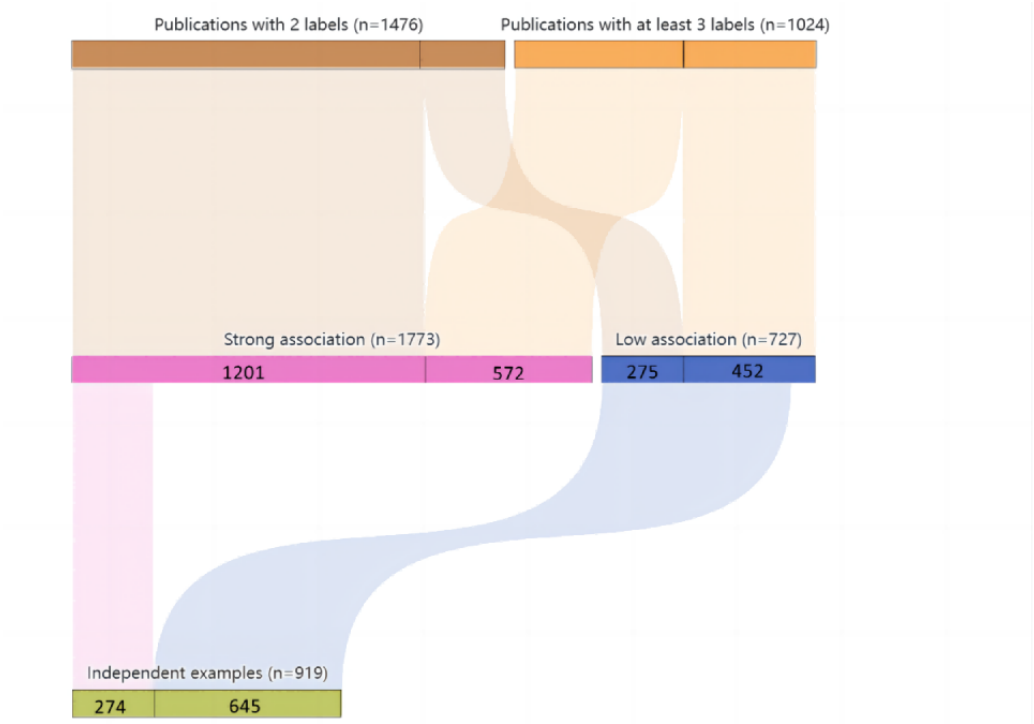


Table 5. Examples of data evaluated by experts.

Sequence number	Title	Labels, n	Correlation strength[a]
1	Temporal Trends of Subsequent Breast Cancer Among Women With Ovarian Cancer: A Population-Based Study [54]	2	1
2	Clinical Characteristics and Survival Outcomes of Patients With Both Primary Breast Cancer and Primary Ovarian Cancer [55]	2	1
3	Secondary Malignancies in Long-Term Ovarian Cancer Survivors: Results of the “Carolin Meets HANNA” Study [56]	2	1
4	Trends in Participation Rates of the National Cancer Screening Program among Cancer Survivors in Korea [57]	3	0
5	Increasing Trends in the Prevalence of Prior Cancer in Newly Diagnosed Lung, Stomach, Colorectal, Breast, Cervical, and Corpus Uterine Cancer Patients: A Population-Based Study [58]	4	1
6	Cancer Registration in China and Its Role in Cancer Prevention and Control [59]	3	0
7	Cancer Incidence, Mortality, and Burden in China: A Time‐Trend Analysis and Comparison With the United States and United Kingdom Based on the Global Epidemiological Data Released in 2020 [60]	5	0
8	Excess Costs and Economic Burden of Obesity-Related Cancers in the United States [61]	3	1
9	Cancer Attributable to Human Papillomavirus Infection in China: Burden and Trends [62]	4	0
10	Excess Costs and Economic Burden of Obesity-Related Cancers in the United States [63]	3	1
11	Cancer Awareness in the General Population varies With Sex, Age and Media Coverage: A Population-Based Survey With Focus on Gynecologic Cancers [64]	5	0
12	Public Attitudes Towards Cancer Survivors Among Korean Adults [65]	3	0
13	Importance of Hospital Cancer Registries in Africa [66]	2	0
14	Correlation Between Family History and Characteristics of Breast Cancer [67]	2	1
15	Familial Aggregation of Early‐Onset Cancers [68]	3	0
16	Trends in Regional Cancer Mortality in Taiwan 1992–2014 [69]	6	0
17	Statin Use and Incidence and Mortality of Breast and Gynecology Cancer: A Cohort Study Using the National Health Insurance Claims database [70]	4	1
18	Management of Breast Cancer Risk in BRCA1/2 Mutation Carriers Who Are Unaffected With Cancer [71]	2	1
19	Association Between Diabetes, Obesity, Aging, and Cancer: Review of Recent Literature [72]	4	0
20	The European Cancer Burden in 2020: Incidence and Mortality Estimates for 40 Countries and 25 Major Cancers [73]	3	0

[a]In the last column, “1” refers to a strong correlation, which means 2 labels of a given publication are semantically or syntactically linked to each other, such as relaying, concurrency, and coupling effects. Conversely, “0” indicates a weak association between multiple labels of a specific publication, and there is no obvious semantic or syntactic correlation.

Figure 7. Relational mapping of multilabel publication distributions.



Discussion

Principal Findings

There are several reasons for the "BERT + TextRNN" model to show optimal performance in cancer publication classification. First, cancer publications usually consist of long texts (eg, titles and abstracts) containing specialty terms and intensive contextual semantic correlations, which are quite suitable for the TextRNN model, which is good at processing sequential information with strong correlation and a high degree of uniformity. Moreover, comprehensive analysis of multilabel classification reflects that cancer publications are characterized by a high multilabel rate in areas with low research intensity due to the interdisciplinary and cooperative working, which enhances the contextual correlation to a certain extent. The "BERT + TextRNN" model is more likely to be efficient in such fields because it can effectively capture contextual semantics.

Compared with the TextRNN, the other models were insufficient and could be further improvement. The TextCNN might not be able to capture sufficient features, since it is not highly interpretable and well suited to address the fixed-length horizon issue. Although the DRNN is an enhanced version of the RNN with low computational speed, it fails to consider any upcoming input to the current state. Therefore, the DRNN is much less effective than the TextRNN. Being a long-linear model, FastText hardly handles the recognition of the long text of cancer publications and needs further optimization due to a limited recall rate.

Limitations

The proposed classifier based on the "BERT + TextRNN" model has 2 issues. On the one hand, the performance of the classifier may be reduced due to the accumulation of errors caused by keyword extraction, which will be enhanced by adjusting the model parameters and adding a self-testing function. On the other hand, the tuple input of titles and abstracts was integrated to train the multilabel classifier, which proved to be better than the input of titles or abstracts alone. Therefore, cancer publications with both title and abstract are desired for the proposed classifier. However, for the few cancer publications papers without abstracts, the classifier we trained will still be usable and has a slight performance cost.

Major Applications of the Proposed Classifier

We trained a classifier based on the "BERT + TextRNN" model for classifying the cancer literature at the publication level, which could directly assign multiple labels to each publication. The proposed classifier has at least 2 major applications. First, the desired model can achieve efficient and effective multilabel classification of cancer publications more granularly, not only for cancer publications in English, but also for full-text literature in other language whose titles and abstracts have English versions. Since the trained classifier is based on cancer publications with titles and abstracts, it should be suitable for any papers whose titles and abstracts are written in English (eg, Chinese medical publications). Another significant application is the fine-grained classification of scientific data on cancer research. Given that valuable data are accompanied by a brief description in English, the proposed model will classify them into the groups with appropriate CTs. Therefore, a content-based label in terms of CTs will be assigned to scientific data and literature, which provides a way to construct a full spectrum of data foundation for precision medicine.

Conclusion

Given that existing classification methods are at the journal level and there is an urgent need for subject classification due to the proliferation of cancer research, a multilabel classifier

was trained based on deep learning models, specifically "BERT + TextRNN." Moreover, the proposed high-resolution classification model was evaluated as being efficient and effective for cancer publications in terms of quantitative comparison and feature analysis.

The innovative exploration in this study is as follows:

- The "BERT + TextRNN" classification model was trained for classifying cancer literature at the publication level, which shows promise in automatically assigning each publication at least 1 label to which it belongs.
- The proposed model achieves high-quality multilabel classification at the publication level, which could reflect the features of cancer publications more accurately with multiple labels compared to the existing method that annotates papers with a single label at the journal level.
- By comprehensive analysis of the correlation between multiple labels, as well as the data characteristics of multilabel cancer publications, the proposed model was verified to be suitable for the literature with features such as high specialization, uniform entity nouns, and standardized long texts.

In the future, the classification model will be extended to classify medical literature on cardiovascular disease and diabetes, where a great number of highly specialized publications have accumulated and are attracting increasing research attention, in order to improve health conditions worldwide.

Acknowledgments

This work was supported by the Innovation Fund for Medical Sciences of Chinese Academy of Medical Sciences (grant: 2021-I2M-1-033) and the Fundamental Research Funds for the Central Universities (grant:3332023163).

Conflicts of Interest

None declared.

References

1. Ferlay J, Ervik M, Lam F, Colombet M, Mery L, Piñeros M. Cancer today: data visualization tools for exploring the global cancer burden in 2020. International Agency for Research on Cancer. 2020. URL: https://gco.iarc.fr/today, [accessed 2023-09-07]
2. Becher T. The significance of disciplinary differences. Stud Higher Educ 1994 Jan;19(2):151-161. [doi: 10.1080/03075079412331382007]
3. Bensaude-Vincent B. Discipline-building in synthetic biology. Stud Hist Philos Biol Biomed Sci 2013 Jun;44(2):122-129 [FREE Full text] [doi: 10.1016/j.shpsc.2013.03.007] [Medline: 23566941]
4. Minaee S, Kalchbrenner N, Cambria E, Nikzad N, Chenaghlu M, Gao J. Deep learning-based text classification: a comprehensive review. ACM Comput Surv 2021 Apr 17;54(3):1-40. [doi: 10.1145/3439726]
5. Wang G, Li C, Wang W, Zhang Y, Shen D, Zhang X, et al. Joint embedding of words and labels for text classification. 2018 Presented at: 56th Annual Meeting of the Association for Computational Linguistics; July 15-20, 2018; Melbourne, Australia. [doi: 10.18653/v1/P18-1216]
6. Conneau A, Schwenk H, Barrault L, Lecun Y. Very deep convolutional networks for text classification. 2017 Presented at: 15th Conference of the European Chapter of the Association for Computational Linguistics; April 2017; Valencia, Spain p. 1107-1116. [doi: 10.18653/v1/e17-1104]
7. Luo P, Wang Y, Wang W. Automatic discipline classification for scientific papers based on a deep pre-training language model. J China Soc Sci Tech Inf 2020;39(10):1046-1059. [doi: 10.1117/12.2311282]
8. CSSC category to Web of Science category mapping 2012. Clarivate Analytics. URL: http://incites.help.clarivate.com/Content/Research%20Areas/china-scadc-subject-categories.htm [accessed 2023-09-09]
9. Hong L, Jia YX. International influence evaluation on disciplines of universities in Shanghai: a bibliometric analysis based on InCites database. Fudan Education Forum 2014;12(4):29-34. [doi: 10.13397/j.cnki.fef.2014.04.006]
10. Fang H. Classifying research articles in multidisciplinary sciences journals into subject categories. KO 2015;42(3):139-153. [doi: 10.5771/0943-7444-2015-3-139]
11. Taheriyan M. Subject classification of research papers based on interrelationships analysis. 2011 Presented at: KDMS 2011: Workshop on Knowledge Discovery, Modeling and Simulation; August 21-24, 2011; San Diego, CA p. 39-44. [doi: 10.1145/2023568.2023579]
12. Wang L, Chen Z, Lin N, Huang X. An interdisciplinary literature classifier based on multi-task multi-label learning. 2021 Presented at: 2021 International Conference on Asian Language Processing (IALP); December 11-13, 2021; Singapore. [doi: 10.1109/ialp54817.2021.9675234]
13. Yan S. Analysis on the evolution trend and characteristics of multidisciplinary oncology team. Shanxi Medical University. 2021. URL: http://kns.cnki.net/kcms/detail/detail.aspx?doi=10.27288/d.cnki.gsxyu.2021.000023&dbcode=CDFD [accessed 2023-09-09]
14. Lyu Q, Feng Y, Shi R, Wu J, Han W. Cancer research driven by cutting-edge multi-disciplinary technologies. Chin Sci Bull 2020 Jul 1;65(31):3446-3460. [doi: 10.1360/tb-2020-0638]

15. Li H, Zheng Z, Xiang F, Wu J, Tan T. Research progress of text classification technology based on deep learning. Comput Eng 2021;47(02):1-11. [doi: 10.19678/j.issn.1000-3428.0059099]
16. Maron ME. Automatic indexing: an experimental inquiry. ACM 1961 Jul;8(3):404-417. [doi: 10.1145/321075.321084]
17. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. J Med Internet Res 2016 Dec 16;18(12):e323 [FREE Full text] [doi: 10.2196/jmir.5870] [Medline: 27986644]
18. Joachims T. Text category with support vector machines: learning with many relevant features. 1998 Presented at: ECML 1998: 10th European Conference on Machine Learning; April 21-23, 1998; Chemnitz, Germany p. 137-142. [doi: 10.1007/bfb0026683]
19. Rasjid ZE, Setiawan R. Performance comparison and optimization of text document classification using k-NN and naïve Bayes classification techniques. Procedia Comput Sci 2017;116:107-112. [doi: 10.1016/j.procs.2017.10.017]
20. Chen L, Tokuda N, Nagai A. A new differential LSI space-based probabilistic document classifier. Inf Process Lett 2003 Dec;88(5):203-212. [doi: 10.1016/j.ipl.2003.09.002]
21. Pennington J, Socher R, Manning CD. GloVe: global vectors for word representation. 2014 Presented at: 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); October 2014; Doha, Qatar. [doi: 10.3115/v1/d14-1162]
22. Mikolov T, Corrado GS, Chen K, Dean J. Efficient estimation of word representations in vector space. arXiv:1301.3781v3 [cs.CL] Preprint posted online January 16, 2013. [FREE Full text] [doi: 10.48550/arXiv.1301.3781]
23. Siwei L. Word and document embeddings based on neural network approaches. arXiv:1611.05962v1 [cs.CL] Preprint posted online November 18, 2016. [FREE Full text] [doi: 10.48550/arXiv.1611.05962]
24. Lai S, Xu L, Liu K, Zhao J. Recurrent convolutional neural networks for text classification. 2015 Presented at: AAAI'15: Twenty-Ninth AAAI Conference on Artificial Intelligence; January 25-30, 2015; Austin, TX. [doi: 10.1609/aaai.v29i1.9513]
25. Kalchbrenner N, Grefenstette E, Blunsom P. A convolutional neural network for modelling sentences. 2014 Presented at: 52nd Annual Meeting of the Association for Computational Linguistics; June 22-27, 2014; Baltimore, MD. [doi: 10.3115/v1/p14-1062]
26. Kim Y. Convolutional neural networks for sentence classificationC. 2014 Presented at: 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); October 2014; Doha, Qatar p. 1746-1751. [doi: 10.3115/v1/d14-1181]
27. Zhang X, Zhao JB, Lecun Y. Character-level convolutional networks for text classificationC. 2015 Presented at: 28th International Conference on Neural Information Processing Systems; December 7-12, 2015; Montreal Canada p. 649-657.
28. Joulin A, Grave E, Bojanowski P, Mikolov T. Bag of tricks for efficient text classification. 2017 Presented at: EACL 2017: 15th Conference of the European Chapter of the Association for Computational Linguistics; April 3-7, 2017; Valencia, Spain p. 427-431. [doi: 10.18653/v1/e17-2068]
29. Hong S, Oh J, Lee H, Han B. Learning transferrable knowledge for semantic segmentation with deep convolutional neural network. 2016 Presented at: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); June 26-July 1, 2016; Las Vegas, NV p. 3204-3212. [doi: 10.1109/cvpr.2016.349]
30. Johnson R, Zhang T. Deep pyramid convolutional neural networks for text categorization. 2017 Presented at: 55th Annual Meeting of the Association for Computational Linguistics; July 2017; Vancouver, Canada p. 562-570. [doi: 10.18653/v1/p17-1052]
31. Xiao Y, Cho K. Efficient character-level document classification by combining convolution and recurrent layers. arXiv:1602.00367v1 [cs.CL] Preprint posted online February 1, 2016. [FREE Full text] [doi: 10.48550/arXiv.1602.00367]
32. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. arXiv:1409.0473v7 [cs.CL] Preprint posted online September 1, 2014. [FREE Full text] [doi: 10.48550/arXiv.1409.0473]
33. Cho K, van Merrienboer B, Gulcehre C, Bougares F, Schwenk H, Bengio Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv:1406.1078v3 [cs.CL] Preprint posted online June 3, 2014. [FREE Full text] [doi: 10.48550/arXiv.1406.1078]
34. Hermann K, Kočiský T, Grefenstette E, Espeholt L, Kay W, Suleyman M, et al. Teaching machines to read and comprehend. arXiv:1506.03340v3 [cs.CL] Preprint posted online June 10, 2015. [FREE Full text] [doi: 10.48550/arXiv.1506.03340]
35. Yang Z, Yang D, Dyer C, He X, Smola A, Hovy E. Hierarchical attention networks for document classification. 2016 Presented at: 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; June 2016; San Diego, CA p. 1480-1489. [doi: 10.18653/v1/n16-1174]
36. Lili S, Peng J, Jing W. A study on the automatic classification of chinese literature in periodicals based on BERT model. Library J 2022;41(05):135-118+135. [doi: 10.13663/j.cnki.lj.2022.05.014]
37. Xiaolei L, Bin N. BERT-CNN: a hierarchical patent classifier based on pre-trained language model. arXiv:1911.06241v1 [cs.CL] Preprint posted online November 3, 2019. [FREE Full text] [doi: 10.48550/arXiv.1911.06241]
38. Liu H, Lin L, Liu C, He H, Wu N, Shen S, et al. Study on the discipline classification of massive humanities and social science academic literature driven by deep learning. Inf Stud Theory Appl 2020 Oct 20:71-81 [FREE Full text] [doi: 10.16353/j.cnki.1000-7490.2023.02.009]
39. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020 Feb 15;36(4):1234-1240 [FREE Full text] [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]

40. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, et al. Domain-specific language model pretraining for biomedical natural language processing. ACM Trans Comput Healthc 2021 Oct 15;3(1):1-23 [FREE Full text] [doi: 10.1145/3458754]
41. Mitra A, Rawat BPS, McManus DD, Yu H. Relation classification for bleeding events from electronic health records using deep learning systems: an empirical study. JMIR Med Inform 2021 Jul 02;9(7):e27527 [FREE Full text] [doi: 10.2196/27527] [Medline: 34255697]
42. Li F, Jin Y, Liu W, Rawat BPS, Cai P, Yu H. Fine-tuning Bidirectional Encoder Representations From Transformers (BERT)–based models on large-scale electronic health record notes: an empirical study. JMIR Med Inform 2019 Sep 12;7(3):e14830 [FREE Full text] [doi: 10.2196/14830] [Medline: 31516126]
43. Chen P, Wang S, Liao W, Kuo L, Chen K, Lin Y, et al. Automatic ICD-10 coding and training system: deep neural network based on supervised learning. JMIR Med Inform 2021 Aug 31;9(8):e23230 [FREE Full text] [doi: 10.2196/23230] [Medline: 34463639]
44. Chen P, Chen K, Liao W, Lai F, He T, Lin S, et al. Automatic International Classification of Diseases coding system: deep contextualized language model with rule-based approaches. JMIR Med Inform 2022 Jun 29;10(6):e37557 [FREE Full text] [doi: 10.2196/37557] [Medline: 35767353]
45. He H, Xia R. Joint binary neural network for multi-label learning with applications to emotion classification. In: Zhang M, Ng V, Zhao D, Li S, Zan H, editors. Natural Language Processing and Chinese Computing. NLPCC 2018. Lecture Notes in Computer Science, Vol 11108. Cham: Springer; 2018.
46. Du J, Chen Q, Peng Y, Xiang Y, Tao C, Lu Z. ML-Net: multi-label classification of biomedical texts with deep neural networks. J Am Med Inform Assoc 2019 Nov 01;26(11):1279-1285 [FREE Full text] [doi: 10.1093/jamia/ocz085] [Medline: 31233120]
47. Glinka K, Woźniak R, Zakrzewska D. Multi-label medical text classification by feature selection. 2017 Presented at: IEEE 26th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE); June 21-23, 2017; Poznan, Poland p. 176-181. [doi: 10.1109/wetice.2017.42]
48. Hughes M, Li I, Kotoulas S, Suzumura T. Medical text classification using convolutional neural networks. arXiv:1704.06841v1 [cs.CL] Preprint posted online April 22, 2017. [FREE Full text] [doi: 10.48550/arXiv.1704.06841]
49. Yogarajan V, Montiel J, Smith T, Pfahringer B. Seeing the whole patient: using multi-label medical text classification techniques to enhance predictions of medical codes. arXiv:2004.00430v1 [cs.IR] Preprint posted online March 29, 2020. [FREE Full text]
50. Wasim M, Asim MN, Ghani Khan MU, Mahmood W. Multi-label biomedical question classification for lexical answer type prediction. J Biomed Inform 2019 May;93:103143 [FREE Full text] [doi: 10.1016/j.jbi.2019.103143] [Medline: 30872137]
51. Baumel T, Nassour-Kassis J, Cohen R, Elhadad M, Elhadad N. Multi-label classification of patient notes: case study on ICD code assignment. arXiv:1709.09587v3 [cs.CL] Preprint posted online September 27, 2012. [FREE Full text]
52. Cancer type list. International Cancer Research Partnership. URL: https://www.icrpartnership.org/cancer-type-list [accessed 2023-09-09]
53. Bryant RE, O'Hallaron D. Computer Systems: A Programmer's Perspective (3rd Edition). London, UK: Pearson; 2016.
54. Matsuo K, Mandelbaum RS, Machida H, Yoshihara K, Matsuzaki S, Klar M, et al. Temporal trends of subsequent breast cancer among women with ovarian cancer: a population-based study. Arch Gynecol Obstet 2020 May;301(5):1235-1245 [FREE Full text] [doi: 10.1007/s00404-020-05508-3] [Medline: 32206877]
55. Chen C, Xu Y, Huang X, Mao F, Shen S, Xu Y, et al. Clinical characteristics and survival outcomes of patients with both primary breast cancer and primary ovarian cancer. Medicine (Baltimore) 2020 Aug 07;99(32):e21560 [FREE Full text] [doi: 10.1097/MD.0000000000021560] [Medline: 32769897]
56. Woopen H, Rolf C, Braicu EI, Buttmann-Schweiger N, Barnes B, Baum J, et al. Secondary malignancies in long-term ovarian cancer survivors: results of the 'Carolin meets HANNA' study. Int J Gynecol Cancer 2021 May 01;31(5):709-712. [doi: 10.1136/ijgc-2020-002155] [Medline: 33649156]
57. Yun EH, Hong S, Her EY, Park B, Suh M, Choi KS, et al. Trends in Participation Rates of the National Cancer Screening Program among Cancer Survivors in Korea. Cancers (Basel) 2020 Dec 30;13(1) [FREE Full text] [doi: 10.3390/cancers13010081] [Medline: 33396692]
58. Sato A, Matsubayashi K, Morishima T, Nakata K, Kawakami K, Miyashiro I. Increasing trends in the prevalence of prior cancer in newly diagnosed lung, stomach, colorectal, breast, cervical, and corpus uterine cancer patients: a population-based study. BMC Cancer 2021 Mar 10;21(1):264 [FREE Full text] [doi: 10.1186/s12885-021-08011-3] [Medline: 33691661]
59. Wei W, Zeng H, Zheng R, Zhang S, An L, Chen R, et al. Cancer registration in China and its role in cancer prevention and control. The Lancet Oncology 2020 Jul;21(7):e342-e349. [doi: 10.1016/s1470-2045(20)30073-5]
60. Qiu H, Cao S, Xu R. Cancer incidence, mortality, and burden in China: a time-trend analysis and comparison with the United States and United Kingdom based on the global epidemiological data released in 2020. Cancer Commun (Lond) 2021 Oct;41(10):1037-1048 [FREE Full text] [doi: 10.1002/cac2.12197] [Medline: 34288593]
61. Hong YR, Huo J, Desai R, Cardel M, Deshmukh AA. Excess Costs and Economic Burden of Obesity-Related Cancers in the United States. Value Health 2019 Dec;22(12):1378-1386 [FREE Full text] [doi: 10.1016/j.jval.2019.07.004] [Medline: 31806194]

62. Lu Y, Li P, Luo G, Liu D, Zou H. Cancer attributable to human papillomavirus infection in China: Burden and trends. Cancer 2020 Aug 15;126(16):3719-3732 [FREE Full text] [doi: 10.1002/cncr.32986] [Medline: 32484937]
63. Hong YR, Huo J, Desai R, Cardel M, Deshmukh AA. Excess Costs and Economic Burden of Obesity-Related Cancers in the United States. Value Health 2019 Dec;22(12):1378-1386 [FREE Full text] [doi: 10.1016/j.jval.2019.07.004] [Medline: 31806194]
64. Fonnes T, Telle IO, Forsse D, Falck R, Trovik J, Haldorsen IS, et al. Cancer awareness in the general population varies with sex, age and media coverage: A population-based survey with focus on gynecologic cancers. Eur J Obstet Gynecol Reprod Biol 2021 Jan;256:25-31 [FREE Full text] [doi: 10.1016/j.ejogrb.2020.10.051] [Medline: 33161211]
65. Kye SY, Lee HJ, Lee Y, Kim YA. Public Attitudes towards Cancer Survivors among Korean Adults. Cancer Res Treat 2020 Jul;52(3):722-729 [FREE Full text] [doi: 10.4143/crt.2019.265] [Medline: 32054152]
66. Curado MP. Importance of hospital cancer registries in Africa. Ecancermedicalscience 2019;13:948 [FREE Full text] [doi: 10.3332/ecancer.2019.948] [Medline: 31552121]
67. Liu L, Hao X, Song Z, Zhi X, Zhang S, Zhang J. Correlation between family history and characteristics of breast cancer. Sci Rep 2021 Mar 18;11(1):6360 [FREE Full text] [doi: 10.1038/s41598-021-85899-8] [Medline: 33737705]
68. Rantala JNJ, Heikkinen SMM, Hirvonen EM, Tanskanen T, Malila NK, Pitkäniemi JM. Familial aggregation of early-onset cancers in early-onset breast cancer families. Int J Cancer 2023 Jul 15;153(2):331-340. [doi: 10.1002/ijc.34538] [Medline: 37074269]
69. Ho YR, Ma SP, Chang KY. Trends in regional cancer mortality in Taiwan 1992-2014. Cancer Epidemiol 2019 Apr;59:185-192. [doi: 10.1016/j.canep.2019.02.005] [Medline: 30825841]
70. Kim DS, Ahn HS, Kim HJ. Statin use and incidence and mortality of breast and gynecology cancer: A cohort study using the National Health Insurance claims database. Int J Cancer 2022 Apr 01;150(7):1156-1165 [FREE Full text] [doi: 10.1002/ijc.33869] [Medline: 34751444]
71. Collins JM, Isaacs C. Management of breast cancer risk in BRCA1/2 mutation carriers who are unaffected with cancer. Breast J 2020 Aug;26(8):1520-1527. [doi: 10.1111/tbj.13970] [Medline: 32652823]
72. Qiang JK, Lipscombe LL, Lega IC. Association between diabetes, obesity, aging, and cancer: review of recent literature. Transl Cancer Res 2020 Sep;9(9):5743-5759 [FREE Full text] [doi: 10.21037/tcr.2020.03.14] [Medline: 35117936]
73. Dyba T, Randi G, Bray F, Martos C, Giusti F, Nicholson N, et al. The European cancer burden in 2020: Incidence and mortality estimates for 40 countries and 25 major cancers. Eur J Cancer 2021 Nov;157:308-347 [FREE Full text] [doi: 10.1016/j.ejca.2021.07.039] [Medline: 34560371]

Abbreviations

BERT: Bidirectional Encoder Representation from Transformers
CNN: convolutional neural network
CT: cancer type
DPCNN: deep pyramid convolutional neural network
DRNN: deep recurrent neural network
ICD: International Classification of Diseases
ICD-10: International Classification of Diseases, Tenth Revision
ICRP CT: International Cancer Research Partnership Cancer Type
LSTM: long short-term memory
MLC: multilabel classification
MLTC: multilabel text classification
RNN: recurrent neural network
TextRNN: text recurrent neural network
TextCNN: text convolutional neural network

Edited by A Benis; submitted 07.12.22; peer-reviewed by Z Ben-Miled, L Guo; comments to author 11.01.23; revised version received 07.03.23; accepted 06.09.23; published 05.10.23.

Please cite as:
Zhang Y, Li X, Liu Y, Li A, Yang X, Tang X
A Multilabel Text Classifier of Cancer Literature at the Publication Level: Methods Study of Medical Text Classification
JMIR Med Inform 2023;11:e44892
URL: https://medinform.jmir.org/2023/1/e44892
doi:10.2196/44892
PMID:37796584

Original Paper

Extending cBioPortal for Therapy Recommendation Documentation in Molecular Tumor Boards: Development and Usability Study

Christopher Renner[1], MSc; Niklas Reimer[2,3], MSc; Jan Christoph[1,4], PhD; Hauke Busch[2,3], PhD; Patrick Metzger[5], PhD; Melanie Boerries[5,6], MD, PhD; Arsenij Ustjanzew[7,8], MSc; Dominik Boehm[9], MSc; Philipp Unberath[1], PhD

[1]Chair of Medical Informatics, Friedrich-Alexander University Erlangen-Nuremberg, Erlangen, Germany

[2]Group for Medical Systems Biology, Lübeck Institute of Experimental, Universität zu Lübeck, Lübeck, Germany

[3]Campus Lübeck, University Cancer Center Schleswig-Holstein, University Hospital Schleswig-Holstein, Lübeck, Germany

[4]Junior Research Group (Bio-) Medical Data Science, Faculty of Medicine, Martin-Luther-University Halle-Wittenberg, Halle, Germany

[5]Institute of Medical Bioinformatics and Systems Medicine, University of Freiburg Faculty of Medicine, University Medical Center Freiburg, Freiburg, Germany

[6]Partner Site Freiburg, German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg, Germany

[7]Institute of Medical Biostatistics, Epidemiology and Informatics (IMBEI), University Medical Center of the Johannes Gutenberg University Mainz, Mainz, Germany

[8]University Cancer Center, University Medical Center of the Johannes Gutenberg-University Mainz, Mainz, Germany

[9]Medical Center for Information and Communication Technology, Universitätsklinikum Erlangen, Erlangen, Germany

Corresponding Author:
Philipp Unberath, PhD
Chair of Medical Informatics
Friedrich-Alexander University Erlangen-Nuremberg
Wetterkreuz 15
Erlangen, 91058
Germany
Phone: 49 1733735424
Email: philipp.unberath@fau.de

Abstract

Background: In molecular tumor boards (MTBs), patients with rare or advanced cancers are discussed by a multidisciplinary team of health care professionals. Software support for MTBs is lacking; in particular, tools for preparing and documenting MTB therapy recommendations need to be developed.

Objective: We aimed to implement an extension to cBioPortal to provide a tool for the documentation of therapy recommendations from MTB sessions in a secure and standardized manner. The developed extension should be embedded in the patient view of cBioPortal to enable easy documentation during MTB sessions. The resulting architecture for storing therapy recommendations should be integrable into various hospital information systems.

Methods: On the basis of a requirements analysis and technology analysis for authentication techniques, a prototype was developed and iteratively refined through a user-centered development process. In conclusion, the tool was evaluated via a usability evaluation, including interviews, structured questionnaires, and the System Usability Scale.

Results: The patient view of cBioPortal was extended with a new tab that enables users to document MTB sessions and therapy recommendations. The role-based access control was expanded to allow for a finer distinction among the rights to view, edit, and delete data. The usability evaluation showed overall good usability and a System Usability Scale score of 83.57.

Conclusions: This study demonstrates how cBioPortal can be extended to not only visualize MTB patient data but also be used as a documentation platform for therapy recommendations.

(JMIR Med Inform 2023;11:e50017) doi:10.2196/50017

KEYWORDS

molecular tumor board; documentation platform; usability evaluation; cBioPortal; precision medicine; genomics; health information interoperability; tumor; implementation; cancer; tool; platform; development; precision; use; user-centered

Introduction

Background

In molecular tumor boards (MTBs), clinical and molecular genetic data from patients with cancer, with a focus on those who lack standard treatment options or have rare tumors, are analyzed and discussed by oncologists, pathologists, and bioinformaticians regarding similarities, abnormalities, and possible new findings. The main goal is both to use advanced molecular genetic diagnostics and clinical assessment to provide therapy recommendations and to gather new insights and potential indications for highly personalized and genome-based therapy recommendations. Such MTBs combine research with patient care and are thus increasingly being implemented by oncologists around the world, as initial studies have shown a benefit for overall patient survival [1-3].

Despite the promising opportunities offered by MTBs, they still face various challenges that need to be addressed. These include the structured documentation of molecular data alongside clinical data and the resulting therapy recommendation in accordance with internal hospital guidelines and patient protection guidelines [4]. However, these data protection guidelines are yet to be uniformly designed and thus handled differently in each hospital [5], requiring local implementation of interfaces to patient records and laboratory systems. Although the transition from paper-based solutions to structured tools such as the electronic health record (EHR) is well advanced, patient-specific MTB therapy recommendations are still often designed as unstructured, free-text fields [5]. Hospitals use either self-programmed software solutions or a combination of various text editors and prefabricated forms, as shown by Hinderer et al [5] in 5 German hospitals. In these hospitals, the genomic data are recorded and communicated electronically or via paper, but in all cases, they are documented as free text without a coordinated data structure. However, this unstructured and nonstandardized state leads to the poor traceability of treatment evidence and decisions, making retrospective or follow-up studies, data sharing, and research projects difficult or even impossible.

Therefore, it is of utmost importance to document MTB recommendations and decisions based on molecular diagnostics in a structured and digital manner to standardize therapy recommendations and medical outcomes across clinics and sites with uniform data formats and reporting rules with the aim to improve patient care, as implied by Buechner et al [6].

Within the Medical Informatics in Research and Care in University Medicine (MIRACUM) consortium [7], the open-source platform cBioPortal [8,9], originally designed as a research platform for storing, analyzing, and visualizing omics data, is used to support MTBs.

To reach this goal, the platform is being adapted, extended, and integrated into university hospital networks so that data can be exchanged among the EHR, the laboratory systems, and the extended cBioPortal [10]. To promote structured therapy recommendations, we developed a therapy recommendation documentation module in cBioPortal, which is suitable for web-based use during MTB sessions as well as for preparation for upcoming MTB sessions.

Therapy recommendations are stored using Health Level Seven Fast Healthcare Interoperability Resources (FHIR) in compliance with the implementation guide “Genomics Reporting” and its German counterpart “Molekulargenetischer Befundbericht” by an FHIR-capable service running alongside cBioPortal [11]. This enables interoperability with the hospital information system (HIS). The standard cBioPortal data model is used for the storage of clinical and molecular patient data. During the development of the module, in addition to developing a standardized way to input the data, the focus was particularly on assuring high usability of the interface for the actual users, as this is a crucial point for the acceptance of a new tool and can even impact the operator’s therapeutic decisions, as shown by Bates et al [12]. To enable secure and compliant use of the data in the hospital network, IT security was taken into account from the beginning of the project. This means that the development was based on best practices regarding the chosen protocols and used interfaces. However, the developments in the area of IT security mainly focus on enabling a connection among the newly created user interface, the HIS, and external backends to the authentication mechanism already existing in the open-source platform and introducing role-based access rights to the backends compatible with commonly used identity providers (IdP). Other security-related aspects, such as enforcing encrypted communication in the network, identifying existing attack vectors, and following all legal and organizational guidelines, must be addressed by the responsible parties during integration into the target network of the respective partner sites.

Objectives

The cBioPortal research platform was expanded with a module that (1) is linked to EHRs and HISs, facilitates the digital documentation of therapy recommendations in MTB sessions in a structured manner, and (2) enables digital accessibility for all persons and partner sites involved in research projects. This allows the standardization of documentation processes across partner sites and supports subsequent collaboration between research and patient care. Therefore, the extension had to be compliant with IT security standards to protect patient data. Furthermore, the solution should be largely independent of the specifics of the HIS in use by providing portable interfaces, which ensures that it can be put into operation at additional sites without major migration effort and thus contribute to an expansion of the research network.

The first objective was to demonstrate the extent to which the solution can harmonize the currently used documentation processes for MTBs across sites. The second objective was to engage future users in a user-centered design approach by

applying an iterative feedback process [13-15] to ensure the quality of the module's use and user satisfaction during the MTB workflow. Finally, to further involve users in the development process and to test the success of this approach, a usability evaluation of the module's web interface was conducted.

Methods

User-Centered Design Approach

During the development of the application, a user-centered design approach was used, which consisted of a requirements analysis together with an iterative feedback process including all partner sites, to improve the usability and, thereby, the clinical applicability of the resulting tool. A total of 12 experts from 6 of the 10 partner sites were involved in this process. They covered a wide variety of backgrounds, including oncology, systems medicine, bioinformatics, and IT, and all of them belonged to the circle of future users. Although the exact processes and roles of an MTB vary from site to site, the use scenario of the documentation platform can be roughly divided into 2 roles. First, there is a preprocessing and postprocessing team, which takes care of data maintenance and documentation. Second, there is a specialist team, which looks at the data and discusses the cases in the MTB. However, both roles can sometimes be performed by the same people, depending on the responsibilities in the site. It was ensured that people from both roles were involved in the user-centered design approach. The selection of additional technologies and frameworks required for authentication and ensuring IT security was elaborated by means of a technology analysis and compared with widespread standards and best practices.

In the first step, the experts specified the required data elements and types from a medical perspective. Subsequently, this information was converted into a data structure applicable to therapy recommendations to be used in the prototype implementation for a high-fidelity mock-up of the documentation module. The resulting mock-up was then iteratively refined in 3 feedback sessions with the help of future users. Each session consisted of a short evaluation with the users, in which they reported feedback or problems via unstructured free text. This feedback was converted into change requests, documented in a quality management system, and integrated into the mock-up before the next feedback session commenced.

Prototype

After the completion of the extended user interface, the module was integrated into the front-end codebase of cBioPortal as a new tab in the patient view. For persisting the entered data, it was connected via a representational state transfer (REST) interface to a FhirSpark server that was used as a tightly coupled, secondary backend [11], resulting in the first fully functional prototype of the module. During the development of the front-end application as well as the backend connections and integrations, an agile approach was used, which was largely based on the adaptive software development concept [16]. The state of the prototype was extended in iterative steps, with the developers meeting at regular intervals with the project team and domain experts to evaluate the recent implementations and discuss the priorities and next features to be implemented from the design phase.

To provide a standardized and simplified way to install the whole setup, cBioPortal and the corresponding services were packaged as Docker containers [17] to make the solution independent of the operating system and use the easy deployment and distribution process of these containers. In the first step, the deployment workflow was tested at the University Hospital Erlangen, including the integration of the components with the HIS. In the second step, the tool was distributed to all 9 consortium sites of MIRACUM and 2 external partner sites to demonstrate that connection to the various systems in the hospital network can be established without major effort and is almost independent of the specific HIS in use. This test was accompanied by structured feedback questionnaires to evaluate whether the general installation was successful, the test data could be imported, and the various tools and annotation services were working as expected. The questionnaire was sent to the partner sites together with the documentation and was meant to be completed by the person responsible for setting up the system together with a key user of the application. This feedback was intended to show whether the planned goals for integrating external sites and distributing the modules can be achieved [10].

Usability Evaluation

Finally, a usability evaluation was conducted using the MIRACUM-cBioPortal (version 2021q1 [18]). This evaluation aimed less at evaluating the resulting application in comparison with other existing tools and more at evaluating the application's general usability in contrast to the previously unstructured approach, as a sufficient level of usability, which leads to users' acceptance, is the basis for a later successful adoption into the clinical workflow. The evaluation was set up as a combination of a task analysis with the thinking-aloud technique, as this approach focuses on the user needs and behaviors while they perform the tasks and can facilitate and accelerate further usability enhancements. In addition to the qualitative evaluation, several quantifiable metrics regarding the time and effort users spent on interacting with the system while completing the tasks were established to determine the extent to which the resulting web interface could meet the self-imposed requirements for improving the documentation process. During the test, users were guided through a series of 8 specific tasks designed to evaluate the system's usability and functionality while articulating their intentions and thoughts as they interacted with the web interface. These tasks encompassed actions such as logging in, creating a new MTB entry for a test patient, manually and automatically generating therapy recommendations, adjusting therapy priorities, and securely logging out. The tasks were formulated to closely mimic a typical workflow and test the essential features of the MTB software system. The users' performance on each task, that is, the time it took them to complete the task and the number of errors they made, as well as their verbally articulated insights, was derived into metrics that allow conclusions about the difficulty of the task itself as well as the amount of time the users spent physically interacting and cognitively engaging with the web interface while working on the given task. After the tasks were completed by the

prospective users, their perceived acceptance of and satisfaction with the module and workflow were measured by interviewing them using a structured questionnaire and the widely used System Usability Scale (SUS) [19] to obtain qualitative feedback as well. In these questionnaires, the users were asked about their level of agreement with preformulated statements about the platform (Multimedia Appendix 1) and performed tasks using a 5-level Likert scale, which ranged from "strongly agree" to "strongly disagree," allowing the derivation of an average satisfaction rate. Finally, the users were given the opportunity to share their opinions and additional subjective remarks to obtain further feedback about their acceptance of the developed module. This feedback was used to assess possible areas of improvement in the interface that could further increase the perceived usability.

Ethical Considerations

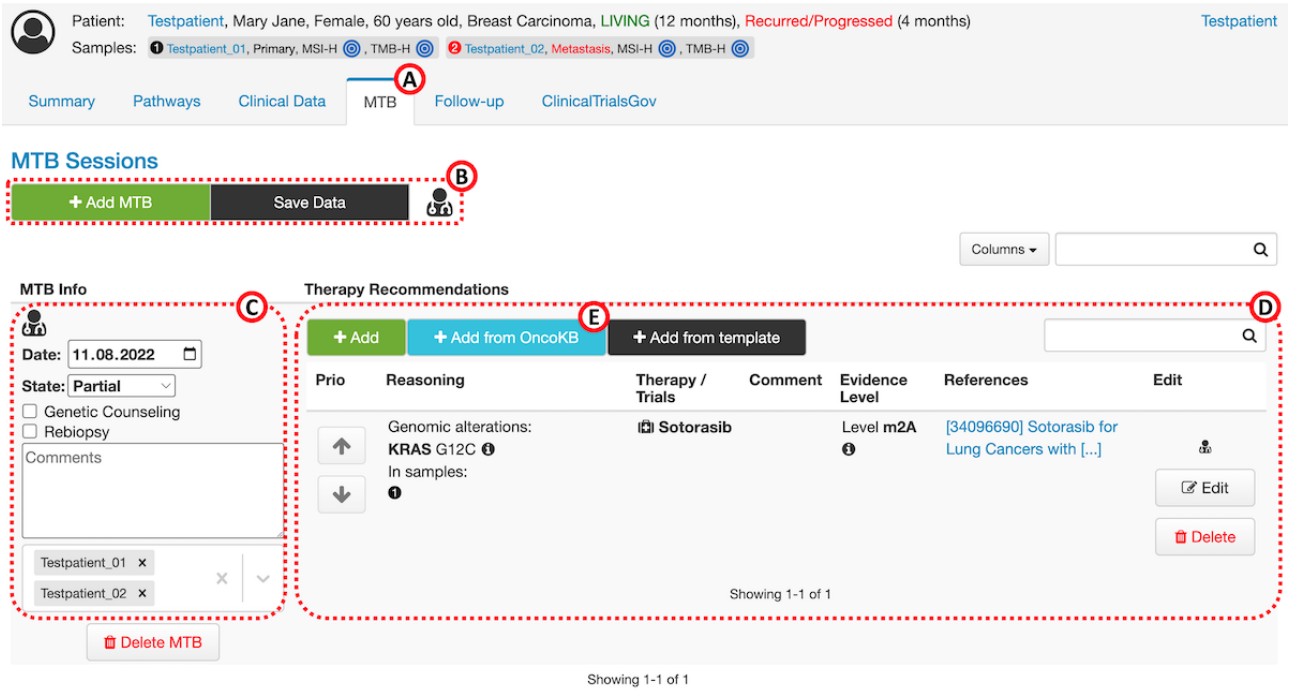
The project was performed in compliance with the World Medical Association Declaration of Helsinki on Ethical Principles for Medical Research Involving Human Subjects. Ethical approval was not required.

Results

User-Centered Design Approach Results

The harmonized data structure of the documentation module in the front end was developed using the iterative approach described earlier. The results showed that in addition to the general and organizational information about the MTB session and the medical data of the discussed patient, further therapy-related characteristics are required. Therefore, the data should include information such as the active substance or substances recommended for therapy, genetic alteration or alterations, complex biomarkers, and clinical data on which the recommendation is based, as well as the level of evidence (defined by the Zentren für Personalisierte Medizin). The clinical and genomic data of the respective patient are already stored in cBioPortal, whereas data about the active substances recommended and level of evidence had to be added and be selectable via a drop-down menu to avoid typing errors and inconsistencies during entry. To substantiate the evidence level, case reports and medical articles listed in the PubMed database [20] can be indicated by their ID, and a field for comments should be available for entering free text. In addition, the knowledge database OncoKB [21], which contains treatment implications of cancer gene alterations, was integrated. A separate button can be used to search for predefined treatment entries in this database and adopt them directly (Figure 1). The mapping of the entries derived from the external database into the module's corresponding input fields is automated, and only the evidence level must be defined manually owing to the use of different evaluation scales. Finally, the design of the web interface was defined, as shown in Figure 1.

Figure 1. Overview of the extended cBioPortal patient view. (A) Newly created tab "MTB" for documenting therapy recommendations. (B) User interface for adding molecular tumor boards (MTBs), saving data, and user management. (C) General information of the MTB session. (D) Detailed documentation of the recommended therapies or trials. (E) Button for adopting matching OncoKB entries as therapy recommendations.



Authentication Enhancement Results

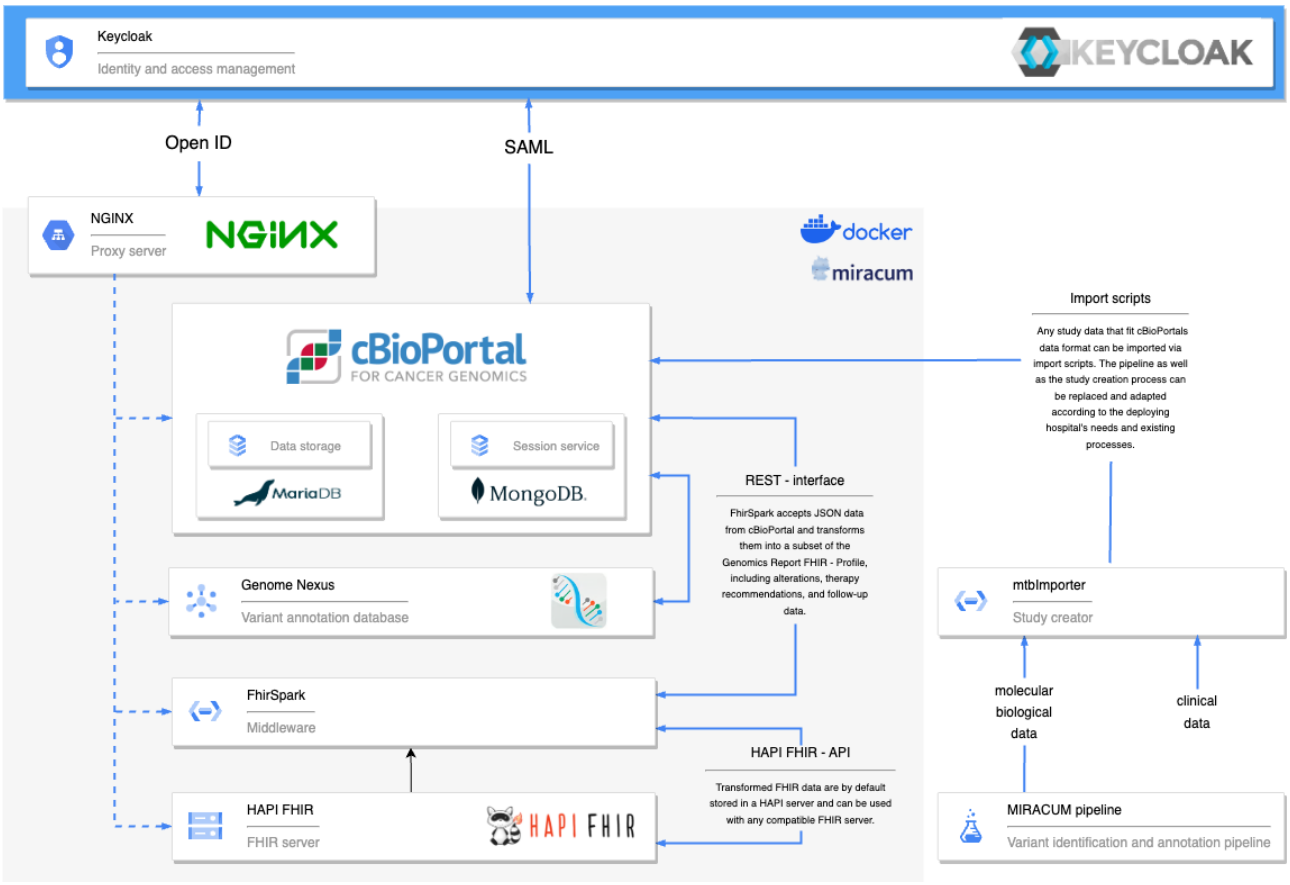
To integrate the module into the HIS and, thereby, comply with the applicable regulations and guidelines for the later planned productive operation, further measures for authentication and authorization had to be implemented in addition to the security features already provided by cBioPortal. Unauthorized access to the data stored in the secondary backend had to be prevented, and a finer distinction between read and write permissions for the data sets had to be enabled.

By default, it is already possible to connect cBioPortal to an external IdP using the Security Assertion Markup Language (SAML) framework [22] to establish a secure connection between the systems and enable individual user log-in sessions. In this process, users can be authorized individually via the IdP for each of the imported studies. For this purpose, the widely used open-source software Keycloak (JBoss) [23] is used, which is supported by Red Hat (Red Hat, Inc) and was already in use at the hospital's network. Keycloak is connected to the corresponding user databases via Lightweight Directory Access Protocol [24], whereby user information can be automatically imported to Keycloak and users can be granted access to cBioPortal, depending on their preassigned roles. However, the therapy documentation module requires the possibility of individually restricting access to patient data as well as a further differentiation of the rights for accessing, editing, and deleting the data. Therefore, it was necessary to set up an additional way of role-based access control for protecting the entered data and to generally restrict access to the FhirSpark backend server. For this purpose, with the lua-resty-openidc module [25], another open-source software was chosen. This module protects the backend by running directly on the nginx proxy server (Nginx, Inc) and can be configured to act as a relying party within the OpenID-Connect [26] authentication layer. For each incoming request, it verifies whether the caller is authenticated by Keycloak and, if so, forwards the user roles attached to the request so that the backend application server can process or deny the request according to the defined logic. Although this architecture uses 2 different protocols, the users are not negatively affected because they only have to log in once via Keycloak. When accessing individual resources, the IdP automatically takes care of the log-in and user role forwarding in the background via single sign-on (SSO). The advantage of outsourcing the access restriction to the nginx proxy server is that with this setup, additional backends can easily be integrated into the SSO realm without them having to deal with authentication protocols, as this is already taken care of by the proxy server.

Prototype Results

For the evaluation of the tool across partner sites, the dockerized workflow was made available via GitHub (GitHub, Inc) [18]. This project combines the extended cBioPortal front end and backend as well as the FhirSpark server, additional annotation services, and further modules for authenticating and authorizing users through a preconfigured nginx proxy server. A detailed system overview is depicted in Figure 2. The tool was deployed and evaluated at 11 partner sites, including the integration of local authentication and authorization services. The study by Reimer et al [10] provides a detailed overview of the evaluation results. Configuring and testing the authentication setup and the subsequent transmission of therapy recommendations to the FhirSpark backend were successful at 4 (36%) of the 11 sites. At the remaining sites (n=7; 64%), the setup was not tested because either a decision was made not to use an authentication service or configuration problems had already occurred at an earlier step, because of which a connection to the secondary backend could not be established. Feedback suggested that the documentation and setup instructions should be extended and adapted to the errors that arose. For example, in addition to listing the minimum hardware requirements and the recommended software versions to rule out errors due to insufficient server configuration, the documentation should be expanded to include an even more detailed description of the Keycloak integration.

Figure 2. Detailed system overview of the extended cBioPortal and the associated services that together form the software platform for supporting molecular tumor boards. API: application programming interface; FHIR: Fast Healthcare Interoperability Resources; HAPI: HL7 Application Programming Interface; MIRACUM: Medical Informatics in Research and Care in University Medicine; REST: representational state transfer; SAML: Security Assertion Markup Language.



Usability Evaluation Results

Among the 12 experts who participated in the user-centered design approach, 7 (58%), all of whom were working in the fields of molecular biology, oncology, and systems medicine and belonging to the group of potential future users, participated in the subsequent usability evaluation conducted to review the extended front-end module. The evaluation showed that, on average, the test users were 89% (SD 19.7%) confident that the enhanced platform can support them with their specific documentation tasks and 71% (SD 36.6%) confident that it can also support the general approach to conducting MTB sessions. With regard to the usability of the web interface itself and the input dialogs, an average satisfaction of 75% (SD 25%) was indicated in terms of the perceived comprehensibility and clarity of the user interface. In addition, the authentication features in particular were rated as satisfactory, with 82% (SD 18.9%) satisfaction in terms of user-friendliness, meaning that the users predominantly did not perceive authentication as disrupting their workflow. The mean score on the administered SUS was 83.6 (SD 14.3). On the basis of the results of the qualitative questionnaire, there is general comprehensibility of the web interface, although in some cases, the correct assignment of clinical data to the designated input fields was rated as not clear enough. For example, some participants attempted to enter relevant biomarkers such as the tumor mutational burden (TMB) value directly into the field for single genomic alterations, possibly thinking that the TMB should be directly assigned to the respective alteration. However, a separate field is provided for this purpose, as values such as the TMB refer not to an alteration but to a specific sample. Apart from providing minor assistance with entering free text so that it gets automatically applied when typing, allowing formatted comments, and more intuitive labeling of the log-in button, no major changes to the user interface were requested. Furthermore, the test users recommended that actual users be given a brief introduction to all the available functionalities at the beginning of productive operation.

For the quantitative evaluation of the task analysis, the proportions of time spent thinking and idling, with thinking and idling combined as cognitive processes, and time spent physically interacting with the web interface while executing the tasks in relation to the total time needed as well as the proportions of time spent for log-in and logout activities were collected. The results are depicted in Figures 3 and 4 and reveal that the actual physical tasks of interacting with the interface and entering new data accounted for the largest share of time spent at 56% (mean 190, SD 70 seconds) of the total time needed, followed by the cognitive processes at 39% (mean 267, SD 101 seconds). A total of 5% of the time was spent making corrections to previously entered data. The process of logging in and out of the platform, as well as verifying the authentication status, accounted for an average of 16% of the total interaction time.

Figure 3. Temporal distribution of the user interaction processes.

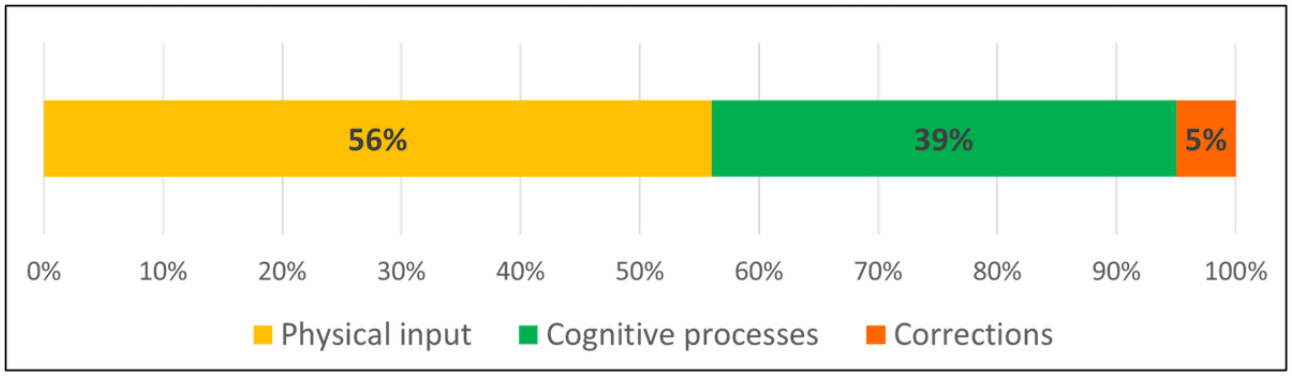
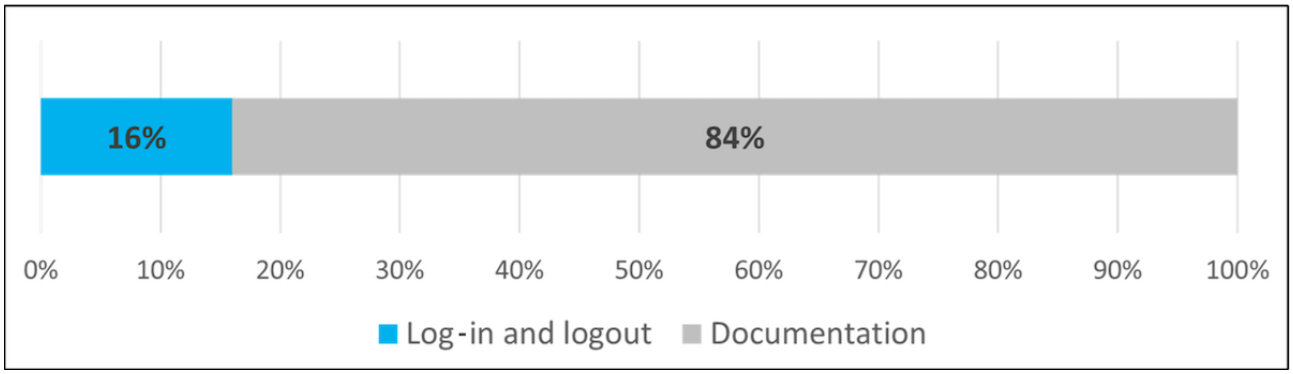


Figure 4. Temporal distribution of the authentication process.



Discussion

User-Centered Design Approach Discussion

A new therapy documentation module for structured therapy recommendations in cBioPortal has been developed using a user-centered design approach and was tested with regard to its functionality and usability. As this was only a prototype implementation, there is still potential for additional enhancements and improvements in usability. The iterative development process revealed its importance through the users' identification of additional requirements at different stages of the prototype. It was found that maximizing the number of feedback rounds with the users provided valuable insights, although the frequency of these rounds was constrained by user availability and the development timeline. Comprehensive documentation was emphasized as crucial for facilitating the deployment phase, allowing for easier adoption and integration of the tool into existing workflows.

Authentication Discussion

To ensure compliance with information security goals, with the SAML framework, a standardized authentication technology is used to protect the connection between the software and the IdP of the hospital. cBioPortal already supports SAML by default; however, no practicable way could be identified to use SAML with the externally developed FhirSpark server, so OpenID-Connect was chosen instead as the security technology for connections to this application. As both protocols are widely used and most IdPs support them, it is possible to flexibly switch from Keycloak to another IdP while maintaining the existing setup and configuration without the need for major adjustments. The use of a proxy server in front of the different backends that takes care of the authentication protocols and forwards the user roles promotes subsequent adaptability to different network environments and the possibility to easily connect additional applications to the same role-based access control system. Furthermore, it is possible within the IdP to enable an SSO session between both protocols and across the several connected backends. With this SSO session, the IdP takes care of the authentication and authorization flow in such a way that users need to log in only once, but their user session is automatically authenticated at all the backends connected via the protocols without losing the advantages and flexibility that come from having 2 security mechanisms in place, with each one suited for its specific purpose. Evidently, certain applications can be excluded from the SSO, and the user permissions can be revoked as necessary. Another possible expansion of the functional scope and, at the same time, a broader differentiation of user permissions for the therapy recommendation module would be the introduction of explicit rights for the deletion of data, which would split up the previously existing editing rights into 2 separate scopes. This would enable the realization of further distinction possibilities between these operations, as there may be use cases where a user has the permission to create or edit entries but not to delete an entire MTB session for traceability and other security reasons. This possible introduction of different rights for the editing and deletion of data was specifically requested by the users; however, in the backend server, there are logging mechanisms that enable the traceability of data records and user actions.

Prototype Feedback

The decision to use a secondary backend for data persistence in the prototype implementation instead of using the existing

internal cBioPortal database is based on the fact that the cBioPortal database is primarily intended as a repository with read-only access to patient data without frequent changes. By contrast, the FhirSpark module is specially designed for changing data sets and provides support for traceability and version control. Furthermore, with each new release of the official cBioPortal backend, the differing database structures would have to be reviewed and potentially merged, which would require an additional maintenance effort. In addition, the proprietary backend offers a higher potential for interface customization and access control, allowing additional external applications to be connected to the database as needed with little effort.

By distributing the prototype implementation to several locations and receiving feedback on the installation and first practical tests, it was determined that the distribution of the application as Docker containers is a suitable method for deploying the software at all partner sites almost independently of the operating systems used and, thereby, accelerating the installation process [10]. This also facilitates the subsequent updates of new versions and changes to individual components. The successful setup and deployment of the platform's basic functionalities were confirmed by all sites, and the newly added components for IT security were successfully connected to the existing solutions and tested at several sites. However, at some sites, problems were encountered at various points in the configuration process, indicating that the documentation needs to be more detailed to cover all the edge cases and specifics of the different IT infrastructures used. To reduce the complexity of the installation process while testing, integrating the tool into the HIS with authentication services was highly recommended but only optional. Although this led to a significant adoption of the tool in general, little feedback was obtained on additional features such as authentication and connection to other systems. Some sites have not implemented these features because of time constraints; others used systems different from Keycloak for identity management, for which configuration instructions were not provided, although integration would be possible. Overall, the feedback provided several indications on how to improve the documentation for future releases as well as further suggestions for minor usability improvements to the user interfaces.

Qualitative Usability Evaluation

Regarding the functionalities offered by the enhanced web interface and the user satisfaction achieved while documenting the therapy recommendation, both indicators were rated as fundamentally positive and satisfactory by the users who participated in the usability evaluation. This shows that the future users are convinced that the module has the potential to digitalize the current processes and documentation data in a structured form and fulfills the general requirements for supporting the existing documentation processes and the expectations of the users. The somewhat lower assessment of the perceived overall support potential for MTB meetings (71%), in contrast to the assessment of the individual functionalities, can be explained by the fact that many sites already use locally established documentation processes, such as their own proprietary implementations or prefabricated text-based document templates. Therefore, converting all workflows to the integrated infrastructure is considered too costly and accepted by stakeholders only with reservations. Another issue is that some users mistakenly related the purpose of the evaluation not only to documentation tasks but also to the whole process of an MTB, starting with case preparation and data analysis. Therefore, the questions were answered from different perspectives. Nevertheless, further feedback should be sought on how to further increase the perceived potential among users. Altogether, the achieved SUS score of 83.6 is above average and can be rated as "excellent" [27] or graded as "A" [28], depending on the evaluation standard applied. Thus, the SUS confirms the results of the structured feedback questionnaire on user satisfaction with and support in using the module. It should be mentioned that the concrete SUS score should not be overinterpreted and can only show a tendency.

Quantitative Usability Evaluation

As the temporal distribution of the processes shows, the users spent most of their time physically entering the data. This is expected because the module is primarily a documentation tool. On the contrary, a predominant share of mental processes would be an alarm sign that the structure of the interface is too complicated. At 39%, the mental processes are nevertheless at a relatively high level, which is due to the fact that most users were using the platform for the first time ever at the time of the evaluation and, therefore, had to first familiarize themselves with the design of the site. In addition, the molecular biology background plays an important role, as the validity of the entries must also be checked during the documentation process and compared with the patient data to avoid errors. The amount of time required to complete the authentication tasks is relatively high (16%). However, this can be explained by the fact that, as part of the tasks, the users had to explicitly check whether the log-in and logout processes were successful and whether they had been granted the correct user rights. After logging out, they also had to actively check whether the therapy recommendations could no longer be modified as part of the test. In the routine workflow, these steps are omitted, and the log-in process needs to be performed only once per session. This will significantly reduce the time spent on authentication tasks. This is also supported by the qualitative feedback from the test users, as the authentication function was subjectively rated as not hindering the workflow and overall perceived as positive.

Furthermore, it can be stated that the participants basically coped well with the operation of the platform but that a short introduction to all the functionalities, user dialogs, and submenus and further test runs before the integration of the platform into the daily clinic routine are considered helpful for quick familiarization. Major changes to the design are not necessary, but the minor suggestions for improving usability that were most frequently mentioned in the feedback, such as the instant adoption of typed text and the sorting of the entered MTB sessions, should be addressed to further improve the quality of use.

Comparison With Other Systems

Although several projects aim to provide an MTB platform and focus on different use cases, none of them yet provide a

standardized means of documenting MTB recommendations and decisions. The Molecular Tumor Board Portal of Tamborero et al [29] offers a clinical decision support platform to analyze Variant Call Format files and generate HTML reports with annotated and interpreted variants. An internal and further expanded version is used by 7 comprehensive cancer centers to analyze sequencing data harmonized with respect to data privacy, state of research, and technological implementation. The platform can be used for joint case discussions, but the documentation thereof is not supported by the portal.

The "VITU – Virtuelles Tumorboard" tool [30] aims to support and digitize MTBs by serving as an information and communication platform. In addition to facilitating MTBs with a videoconferencing tool integrated into the platform and a digitized, process-based case review option, further solutions for integrating external experts will be offered, according to their website. It also contains a module for structured documentation in combination with an FHIR interface for accessing the data. Although there is a demo version of the software available for web-based testing and the source code is available on GitHub, it is not yet conceivable how the software would perform in productive operations, as it still seems to be in the development phase, and updates seem to have slowed down since version 2019.3 [31].

A similar purpose in terms of supporting and promoting MTBs is pursued by Alteration Annotations for Molecular Tumor Boards [32]. This is an R shiny–based web application developed by researchers at the University Hospital in Ulm and consists of multiple modules that, in addition to the visualization and annotation of mutations, offer the possibility of displaying the evidence for possible therapeutic drug targets. Thereby, the identified mutations can be evaluated, discussed, and processed in various formats so that the findings can be exported to external clinical systems. Therefore, the public databases GDKD, CIViC, and TARGET [33-35] can be connected, which are used as a knowledge basis for variant annotations by linking together various information sources. Overall, the application, which is based on the R programming language, is primarily designed for the preparation and visualization of data but also includes functions for highly automated data integration and the standardized documentation of results. However, the capability to share data across sites, specifically for research purposes, appears to be limited.

Schapranow et al [36] have created an advanced software tool designed specifically for multidisciplinary tumor boards. Their approach involved a thorough requirements analysis and followed a user-centered development process. The software allows a dedicated person to note down therapy recommendations, clinical trials, and additional remarks live and transparently on a presentation screen visible to the attendees. The software then automatically generates a report upon the completion of the structured documentation, which is provided to the treating physician. In addition, the software integrates external data sources in a modular manner, focusing on the specific requirements of multidisciplinary tumor boards, and builds the technical foundation of a multisite database of participating hospitals. Although the software is still in the prototype phase, its modular design should enable easy integration with existing clinical documentation.

Comparing the functional scope of cBioPortal in combination with the newly created module with that of similar, already existing systems, it becomes apparent that the integration of a standardized documentation tool for the cross-site harmonization of findings and therapy approaches can create an added value that is not yet fully exploited by other solutions. The possibility of integrating the tool into the clinic's internal workflow and additionally connecting other external systems and data sources to it holds great potential, particularly with regard to the exchange of research data for the transfer of knowledge and the promotion of innovation between clinic alliances and research networks that this makes possible. Thus, there are many more use cases that can be addressed beyond just data visualization and storage, such as fostering collaboration between research and patient care and facilitating the selection and inclusion of appropriate data sets and patients for studies and research projects. That this is a contemporary application field with great potential is demonstrated by the National Network of Genomic Medicine, in which personalized therapy for patients with lung cancer is supported through the provision of a common database and facilitated by the use of a central study registry [37].

Finally, the question of whether the developed solution might be used for similar applications, especially organ tumor boards, arises. Although there is a significant overlap in the general requirements, as well as some overlap in the underlying data model, cBioPortal is primarily tailored to molecular biological data and their visualization. Therefore, it is not an adequate software basis for other tumor board applications without significant alterations. That said, the general approach to design and development is replicable and may be adapted to develop solutions for related problems. One of the key points of this study that can be generalized to the field of medical informatics is that it is crucial to develop tools that adhere to established standards, especially using FHIR and harmonized implementation guides. This ensures seamless integration. In addition, consistently prioritizing user requirements is essential for effective and efficient support for health care delivery.

Outlook

Using the feedback from the usability evaluation, the tool will be refined and adapted to provide users with a comprehensive documentation platform for MTBs. Subsequently, an additional tab will be developed to enable the documentation of follow-up data based on the recommended therapies to further expand the documentation capabilities of cBioPortal. This also paves the way for building a standardized and multisite pool of MTB data, which can later be used for the annotation of new patients and as input for machine learning techniques to generate new insights about MTB treatment and outcomes. An important aspect of extending the tool is compliance with the Medical Device Regulation and In Vitro Diagnostic Device Regulation. As the software currently supports (in the context of Medical Device Coordination Group 2019-11 [38]) only simple functionalities, such as storage, archiving, communication, and simple search, it does not fall within these regulations.

Conclusions

In summary, a custom module for the cBioPortal platform was implemented to enhance the platform with functionalities for documenting therapy recommendations in a harmonized manner across collaborating hospitals and research sites, and an interface for enhanced authentication and authorization purposes was added to comply with security regulations and ensure the protection of patients' data. The findings of the hands-on tests and the usability evaluation suggest that the resulting solution for the documentation of therapy recommendations can be deployed successfully at various sites independently of the HIS in use without much effort. The usability evaluation leads to the assumption that the module and user interface will be widely accepted by the future users and thus can be successfully integrated into their workflow. Therefore, the newly introduced functionalities have the potential to improve the existing documentation processes by providing a structured and harmonized digital template for documentation data. Thus, these functionalities could not only lead to a significant benefit by directly supporting the preparation and conduct of MTBs but also promote the development of further applications, leveraging the implemented harmonized data structure to reuse the collected data. Fields of possible applications are the postprocessing of MTBs and case preparation using the follow-up data of previous patients in future MTBs. In addition, opportunities for interdisciplinary exchange may arise through the transfer of innovation among different research groups to discover new medical relationships based on the accessibility and sharing of data among partner sites.

Acknowledgments

The authors thank all the participants of the usability study for their time and effort.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Statements of the quantitative questionnaire and items of the open feedback questionnaire (translated from German for publishing).
[DOCX File , 15 KB - medinform_v11i1e50017_app1.docx]

References

1. Hoefflin R, Geißler AL, Fritsch R, Claus R, Wehrle J, Metzger P, et al. Personalized clinical decision making through implementation of a molecular tumor board: a German single-center experience. JCO Precis Oncol 2018 Nov;2:1-16. [doi: 10.1200/po.18.00105]
2. Hoefflin R, Lazarou A, Hess ME, Reiser M, Wehrle J, Metzger P, et al. Transitioning the molecular tumor board from proof of concept to clinical routine: a German single-center analysis. Cancers (Basel) 2021 Mar 08;13(5):1151 [FREE Full text] [doi: 10.3390/cancers13051151] [Medline: 33800365]
3. Horak P, Leichsenring J, Kreutzfeldt S, Kazdal D, Teleanu V, Endris V, et al. [Variant interpretation in molecular pathology and oncology : an introduction]. Pathologe 2021 Jul 03;42(4):369-379. [doi: 10.1007/s00292-021-00938-5] [Medline: 33938987]
4. Haier J, Bergmann KO. [Medicolegal aspects of tumor boards]. Chirurg 2013 Mar 1;84(3):225-230. [doi: 10.1007/s00104-013-2481-4] [Medline: 23455588]
5. Hinderer M, Boerries M, Haller F, Wagner S, Sollfrank S, Acker T, et al. Supporting molecular tumor boards in molecular-guided decision-making - the current status of five German university hospitals. Stud Health Technol Inform 2017;236:48-54. [Medline: 28508778]
6. Buechner P, Hinderer M, Unberath P, Metzger P, Boeker M, Acker T, et al. Requirements analysis and specification for a molecular tumor board platform based on cBioPortal. Diagnostics (Basel) 2020 Feb 10;10(2):93 [FREE Full text] [doi: 10.3390/diagnostics10020093] [Medline: 32050609]
7. Prokosch HU, Acker T, Bernarding J, Binder H, Boeker M, Boerries M, et al. MIRACUM: medical informatics in research and care in university medicine. Methods Inf Med 2018 Jul 17;57(S 01):e82-e91. [doi: 10.3414/me17-02-0025]
8. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov 2012 May;2(5):401-404 [FREE Full text] [doi: 10.1158/2159-8290.CD-12-0095] [Medline: 22588877]
9. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal 2013 Apr 02;6(269):pl1 [FREE Full text] [doi: 10.1126/scisignal.2004088] [Medline: 23550210]
10. Reimer N, Unberath P, Busch H, Börries M, Metzger P, Ustjanzew A, et al. Challenges and experiences extending the cBioPortal for cancer genomics to a molecular tumor board platform. Stud Health Technol Inform 2021 Nov 18;287:139-143. [doi: 10.3233/SHTI210833] [Medline: 34795098]

11. Reimer N, Unberath P, Busch H, Ingenerf J. FhirSpark - implementing a mediation layer to bring for FHIR to the cBioPortal cancer genomics. Stud Health Technol Inform 2021 May 27;281:303-307. [doi: 10.3233/SHTI210169] [Medline: 34042754]
12. Bates DW, Kuperman GJ, Wang S, Gandhi T, Kittler A, Volk L, et al. Ten commandments for effective clinical decision support: making the practice of evidence-based medicine a reality. J Am Med Inform Assoc 2003 Nov 01;10(6):523-530. [doi: 10.1197/jamia.m1370]
13. Abras C, Maloney-Krichmar D, Preece J. User-centered design. In: Bainbridge W, editor. Encyclopedia of Human-Computer Interaction. Thousand Oaks, CA: SAGE Publications; 2004.
14. da Silva TS, Martin A, Maurer F, Silveira M. User-centered design and agile methods: a systematic review. In: Proceedings of the 2011 Agile Conference. 2011 Presented at: 2011 Agile Conference; August 07-13, 2011; Salt Lake City, UT URL: https://ieeexplore.ieee.org/document/6005488 [doi: 10.1109/agile.2011.24]
15. Still B, Crane K. Fundamentals of User-Centered Design: A Practical Approach. Boca Raton, FL: CRC Press; 2017.
16. Highsmith J. Adaptive Software Development: A Collaborative Approach to Managing Complex Systems. London, UK: Pearson Education; 2013.
17. Anderson C. Docker [Software engineering]. IEEE Softw 2015 May;32(3):102-105. [doi: 10.1109/ms.2015.62]
18. MIRACUM-cbioportal. GitHub. URL: https://github.com/buschlab/MIRACUM-cbioportal [accessed 2023-10-25]
19. Brooke J. SUS: a 'quick and dirty' usability scale. In: Usability Evaluation In Industry. Boca Raton, FL: CRC Press; 1996.
20. Canese K, Weis S. PubMed: the bibliographic database. The NCBI Handbook. 2002. URL: https://www.ehu.eus/biofisica/juanma/mbb/pdf/pubmed_intro.pdf [accessed 2023-10-25]
21. Chakravarty D, Gao J, Phillips SM, Kundra R, Zhang H, Wang J, et al. OncoKB: a precision oncology knowledge base. JCO Precis Oncol 2017 Jul;2017:PO.17.00011 [FREE Full text] [doi: 10.1200/PO.17.00011] [Medline: 28890946]
22. Security Assertion Markup Language (SAML) V2.0 technical overview. OASIS. 2008. URL: https://www.oasis-open.org/committees/download.php/27819/sstc-saml-tech-overview-2.0-cd-02.pdf [accessed 2023-10-25]
23. Divyabharathi DN, Cholli NG. A review on identity and access management server (KeyCloak). Int J Secur Priv Pervasive Comput 2020;12(3):46-53. [doi: 10.4018/IJSPPC.2020070104]
24. Zeilenga K. Lightweight directory access protocol (LDAP): technical specification road map. Internet Engineering Task Force Datatracker. URL: https://datatracker.ietf.org/doc/rfc4510/ [accessed 2023-10-25]
25. lua-resty-openidc. GitHub. URL: https://github.com/zmartzone/lua-resty-openidc [accessed 2023-10-25]
26. Fett D, Küsters R, Schmitz G. The web SSO standard OpenID connect: in-depth formal security analysis and security guidelines. In: Proceeding of the IEEE 30th Computer Security Foundations Symposium (CSF). 2017 Presented at: IEEE 30th Computer Security Foundations Symposium (CSF); August 21-25, 2017; Santa Barbara, CA. [doi: 10.1109/csf.2017.20]
27. Bangor A, Kortum P, Miller J. Determining what individual SUS scores mean: adding an adjective rating scale. J Usability Stud 2009;4(3):114-123.
28. Lewis JR, Sauro J. Item benchmarks for the system usability scale. J Usability Stud 2018 May 1;13(3):158-167.
29. Tamborero D, Dienstmann R, Rachid MH, Boekel J, Lopez-Fernandez A, Jonsson M, et al. The Molecular Tumor Board Portal supports clinical decisions and automated reporting for precision oncology. Nat Cancer 2022 Feb 24;3(2):251-261 [FREE Full text] [doi: 10.1038/s43018-022-00332-x] [Medline: 35221333]
30. Fegeler C, Zsebedits D, Bochum S, Finkeisen D, Martens UM. Implementierung eines IT-gestützten molekularen tumorboards in der regelversorgung. Forum 2018 Aug 14;33(5):322-328. [doi: 10.1007/s12312-018-0459-3]
31. Einführung VITU 2019.3. MOLIT Docs. URL: https://docs.molit.eu/molit-docs/v2019.3/guide/ [accessed 2022-04-13]
32. Fürstberger A, Ikonomi N, Kestler AM, Marienfeld R, Schwab JD, Kuhn P, et al. AMBAR - interactive alteration annotations for molecular tumor boards. Comput Methods Programs Biomed 2023 Oct;240:107697 [FREE Full text] [doi: 10.1016/j.cmpb.2023.107697] [Medline: 37441893]
33. Dienstmann R, Jang IS, Bot B, Friend S, Guinney J. Database of genomic biomarkers for cancer drugs and clinical targetability in solid tumors. Cancer Discov 2015 Feb;5(2):118-123 [FREE Full text] [doi: 10.1158/2159-8290.CD-14-1118] [Medline: 25656898]
34. Griffith M, Spies NC, Krysiak K, McMichael JF, Coffman AC, Danos AM, et al. CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. Nat Genet 2017 Jan 31;49(2):170-174 [FREE Full text] [doi: 10.1038/ng.3774] [Medline: 28138153]
35. Le Tourneau C, Delord JP, Gonçalves A, Gavoille C, Dubot C, Isambert N, et al. Molecularly targeted therapy based on tumour molecular profiling versus conventional therapy for advanced cancer (SHIVA): a multicentre, open-label, proof-of-concept, randomised, controlled phase 2 trial. Lancet Oncol 2015 Oct;16(13):1324-1334. [doi: 10.1016/S1470-2045(15)00188-6] [Medline: 26342236]
36. Schapranow MP, Borchert F, Bougatf N, Hund H, Eils R. Software-tool support for collaborative, virtual, multi-site molecular tumor boards. SN Comput Sci 2023 Apr 27;4(4):358 [FREE Full text] [doi: 10.1007/s42979-023-01771-8] [Medline: 37131499]
37. Büttner R, Wolf J, Kron A, Nationales Netzwerk Genomische Medizin. [The national Network Genomic Medicine (nNGM) : model for innovative diagnostics and therapy of lung cancer within a public healthcare system]. Pathologe 2019 May 17;40(3):276-280. [doi: 10.1007/s00292-019-0605-4] [Medline: 31101971]

38. Guidance on qualification and classification of software in regulation (EU) 2017/745 – MDR and regulation (EU) 2017/746 – IVDR. European Commission. 2019 Oct 10. URL: https://ec.europa.eu/docsroom/documents/37581 [accessed 2023-10-25]

Abbreviations

EHR: electronic health record
FHIR: Fast Healthcare Interoperability Resources
HIS: hospital information system
MIRACUM: Medical Informatics in Research and Care in University Medicine
MTB: molecular tumor board
REST: representational state transfer
SAML: Security Assertion Markup Language
SSO: single sign-on
SUS: System Usability Scale
TMB: tumor mutational burden

Edited by C Lovis; submitted 16.06.23; peer-reviewed by R Taira, M Schapranow; comments to author 14.07.23; revised version received 02.09.23; accepted 17.09.23; published 11.12.23.

Please cite as:
Renner C, Reimer N, Christoph J, Busch H, Metzger P, Boerries M, Ustjanzew A, Boehm D, Unberath P
Extending cBioPortal for Therapy Recommendation Documentation in Molecular Tumor Boards: Development and Usability Study
JMIR Med Inform 2023;11:e50017
URL: https://medinform.jmir.org/2023/1/e50017
doi:10.2196/50017
PMID:38079196

Diagnostic Performance, Triage Safety, and Usability of a Clinical Decision Support System Within a University Hospital Emergency Department: Algorithm Performance and Usability Study

Juhani Määttä[1], MD, PhD; Rony Lindell[1], MD; Nick Hayward[1], MD, PhD; Susanna Martikainen[2], PhD; Katri Honkanen[3], MD; Matias Inkala[3], MD; Petteri Hirvonen[1], MD; Tero J Martikainen[3], MD, PhD

1

2

3

Corresponding Author:
Tero J Martikainen, MD, PhD

Abstract

Background: Computerized clinical decision support systems (CDSSs) are increasingly adopted in health care to optimize resources and streamline patient flow. However, they often lack scientific validation against standard medical care.

Objective: The purpose of this study was to assess the performance, safety, and usability of a CDSS in a university hospital emergency department setting in Kuopio, Finland.

Methods: Patients entering the emergency department were asked to voluntarily participate in this study. Patients aged 17 years or younger, patients with cognitive impairments, and patients who entered the unit in an ambulance or with the need for immediate care were excluded. Patients completed the CDSS web-based form and usability questionnaire when waiting for the triage nurse's evaluation. The CDSS data were anonymized and did not affect the patients' usual evaluation or treatment. Retrospectively, 2 medical doctors evaluated the urgency of each patient's condition by using the triage nurse's information, and urgent and nonurgent groups were created. The *International Statistical Classification of Diseases, Tenth Revision* diagnoses were collected from the electronic health records. Usability was assessed by using a positive version of the System Usability Scale questionnaire.

Results: In total, our analyses included 248 patients. Regarding urgency, the mean sensitivities were 85% and 19%, respectively, for urgent and nonurgent cases when assessing the performance of CDSS evaluations in comparison to that of physicians. The mean sensitivities were 85% and 35%, respectively, when comparing the evaluations between the two physicians. Our CDSS did not miss any cases that were evaluated to be emergencies by physicians; thus, all emergency cases evaluated by physicians were evaluated as either urgent cases or emergency cases by the CDSS. In differential diagnosis, the CDSS had an exact match accuracy of 45.5% (97/213). The usability was good, with a mean System Usability Scale score of 78.2 (SD 16.8).

Conclusions: In a university hospital emergency department setting with a large real-world population, our CDSS was found to be equally as sensitive in urgent patient cases as physicians and was found to have an acceptable differential diagnosis accuracy, with good usability. These results suggest that this CDSS can be safely assessed further in a real-world setting. A CDSS could accelerate triage by providing patient-provided data in advance of patients' initial consultations and categorize patient cases as urgent and nonurgent cases upon patients' arrival to the emergency department.

Trial Registration: ClinicalTrials.gov NCT04577079; https://www.clinicaltrials.gov/study/NCT04577079

(JMIR Med Inform 2023;11:e46760) doi:10.2196/46760

KEYWORDS

clinical decision support system; emergency department; performance; usability; user experience; validation; medical care; decision-making; digital health; differential diagnosis; triage; patient population

Introduction

Digital health technology is increasingly being developed to support health care systems around the world. Digital health technology includes solutions that assess the urgency and the differential diagnosis of the patient's symptoms with the help of artificial intelligence. These solutions are frequently called *clinical decision support systems* (CDSSs) or *computerized diagnostic decision support programs* [1,2].

These systems can either aid health care professionals in their decision-making or give information on symptoms, conditions, and possible recommendations for future actions to the patient. However, patient-provided data solutions are rarely scientifically validated [3,4]. There are very few studies that have assessed

solution performances in real-world settings, and these are rarely performed using a patient sample with a broad range of conditions [3,5-8].

Work in triage and emergency departments is highly demanding, with substantial time pressure. There is a risk of human errors when operating with high patient volumes, acute conditions, and severe stress. Physicians have been estimated to have a 5% diagnostic error rate [9], with half of these errors being potentially harmful [10]. In an emergency department setting, a CDSS based on patient-provided data that are shared prior to patients entering the emergency department could have the potential to significantly help in allocating optimal resources for the patients who need more prompt assessments and complex care.

The extensive need for the validation of any CDSS with patient-provided data has been acknowledged widely [2,4,5,7]. Studies that evaluate the accuracy or diagnostic performance of a CDSS with patient-provided data in a wide and diverse set of patients are rare and have mainly used clinical vignettes rather than actual patients in a real-life setting [3,4]. A clinical vignette study can describe the experimental accuracy and performance of a system algorithm, but it (1) omits the complex diversities and randomness among patients in real life, (2) typically concentrates only on textbook cases, and (3) omits the usability of the system. Combined clinical vignette studies mainly describe the theoretical performance of an algorithm and do not validate the performance of the CDSS in actual use. Hence, a CDSS with patient-provided data should always take usability into consideration. High usability is one of the key factors that allow a system to make correct interpretations when dealing with patient-provided data. The system needs to be widely easy to use and effective (ie, usable) [11]. One of the most popular questionnaires for assessing system usability is the System Usability Scale (SUS) questionnaire [12-14]. A positive version of the SUS questionnaire has, in addition, been developed [13].

The aim of this study was to analyze the performance of a particular CDSS (Klinik Access [Klinik Healthcare Solutions Oy]) with patient-provided data in a real-world university hospital emergency department setting. The main aims were to (1) evaluate the performance of the system in detecting the clinical urgency of a patient's condition and (2) evaluate diagnostic performance by using the actual *International Statistical Classification of Diseases, Tenth Revision* (*ICD-10*) diagnoses assigned at the emergency department. Our third aim was to assess the usability of the system in a real-world setting with real patients. Our hypotheses were that the CDSS has acceptable safety margins and sensitivity in clinically urgent cases, diagnostic performance correlates well with actual medical diagnoses, and the usability of the system is good or better.

Methods

Study Population

The study population consisted of patients who entered the emergency department of Kuopio University Hospital, Finland, during a 3-week period in September 2020. The patients were recruited by the research assistants in the emergency department waiting room while waiting for a triage nurse's assessment. Participation in this study was fully voluntary. The following patients were excluded from this study: patients aged 17 years or younger, patients with cognitive impairments, and patients who entered the unit in an ambulance or with the need for immediate care. The patients were informed about this study, including the information that this study would not affect their emergency department visit or received treatment in any way. To verify their acceptance to participate in this study, the patients provided written informed consent.

Ethics Approval

This prospective study has been reported in ClinicalTrials.gov (NCT04577079), and this study was approved by the research ethics committee of the Northern Savo Hospital District (347/2020), Finland.

Data Collection, Urgency Evaluation, and Diagnosis

After consenting to study participation, patients completed the CDSS (Klinik Access) web-based form independently, using tablet computers that were provided by the research assistants. The form was used to obtain information, including demographics, history, and symptom-related factors. The data provided by the patients were not used by the emergency department personnel. After the data collection phase, 2 independent physicians (KH, MD [15 y of experience]; MI, MD [7 y of experience]) assessed the patient data from the electronic health records. The assessing physicians did not have any access to or knowledge of the CDSS data.

First, the physicians assessed the urgency of each patient case, using only written information provided by the triage nurse—the identical information on which urgency is based in the emergency department. Urgency was mapped to the same four categories that the CDSS software used (Table 1). After evaluating the urgency, the assessing physicians reviewed the visits in the electronic health records and checked that the correct *ICD-10* diagnoses were provided by the emergency department clinicians.

Table . Evaluation of the urgency of a patient's condition and clinical examples of related conditions.

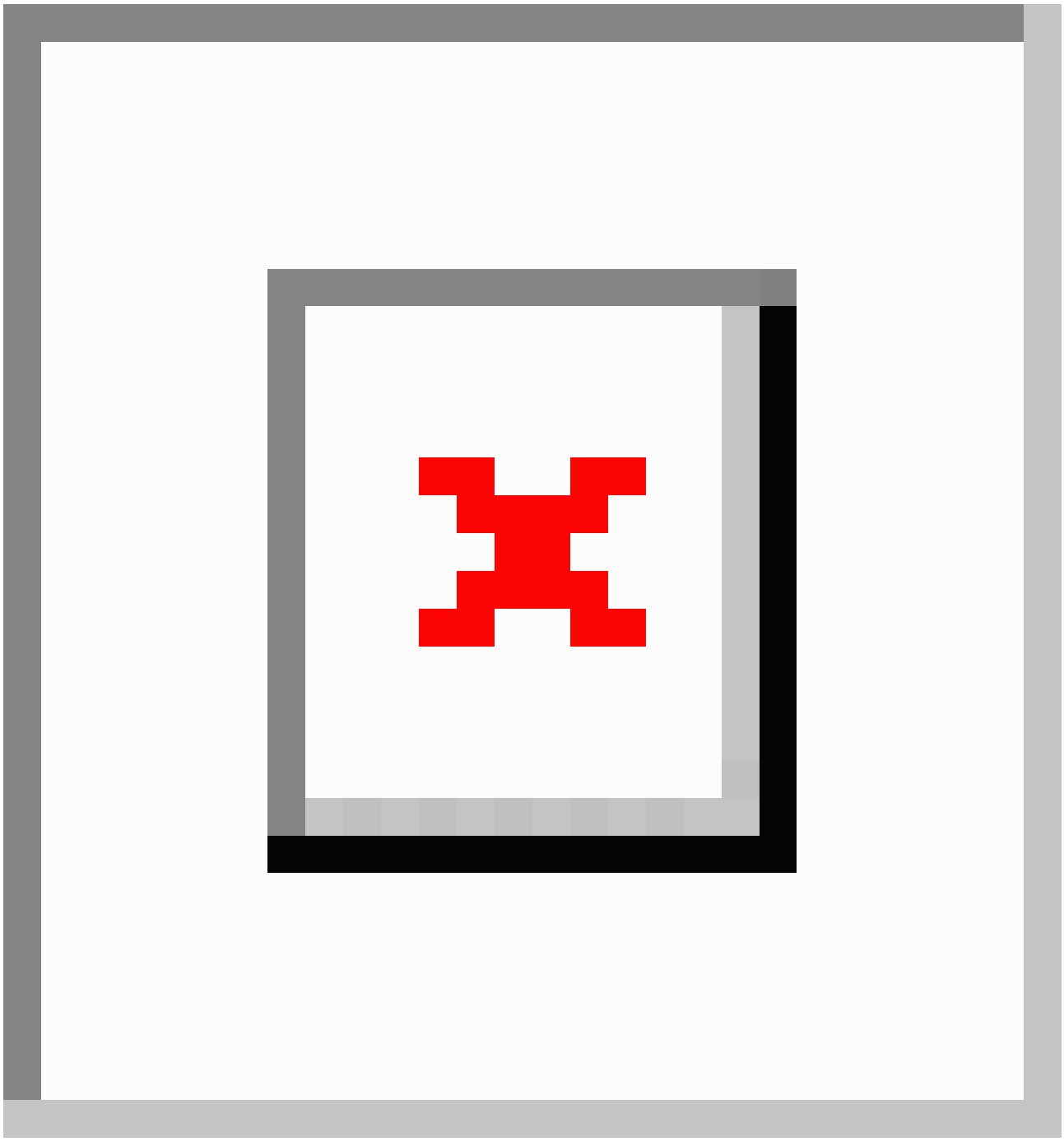
Urgency	Definition	Clinical example (symptom or condition)
Self-care	Benign condition that could primarily be self-treated by the patient (ie, no real need for assessment by a clinician)	• Mild viral infection • Hives (urticaria)
Nonurgent	Condition that requires nonurgent evaluation by a clinician (ie, patient should primarily receive a nonurgent [general practitioner] appointment)	• Prolonged cough in an otherwise healthy young patient • Symptomatic knee arthrosis without significant disability
Urgent	Condition that requires evaluation by a clinician within the next few days, preferably during the same day	• Ear pain with only temporary relief via analgesics • Severe shoulder pain from an injury
Emergency	Condition that should be evaluated by a clinician within 2 hours or as soon as possible	• Chest pain • Breathing difficulty

The CDSS

The CDSS used in this study was Klinik Access (version 1.1). The system intends to support primary care health care organizations by receiving service inquiries from patients, automatically preanalyzing the included clinical information, and supporting health care professionals in managing the triage demand effectively. It has already been implemented in over 400 primary care and dental facilities in the Nordics, the United Kingdom, and the Netherlands. Klinik Access evaluates patient symptoms and potential conditions by using the situational data provided by the patients (Figure 1). It consists of an interface for receiving the patient inputs, a back end for generating the computational dynamic questionnaire and other business logic, and a professional interface for the management of the patient inquiries by health care staff. It was originally developed for general primary care use, including children and adolescent care use. Although it properly recognizes various urgent and emergency situations, it has not yet been optimized for use solely in an emergency department context. As such, the clinical context in this study was intended to be partially experimental.

The Klinik Access system uses a medical gold standard Bayesian methodology [15,16] for inferring the effect of clinical features on the probabilities of the differential diagnosis conditions. The severity and the urgency of a condition were inferred, in addition to the differential diagnosis, by using specific severity symptoms and by setting a threshold (15%) for the probabilities of relevant conditions in the differential. The purpose of this study was to analyze the performance of the algorithm and the patient usability of the CDSS. As such, only the system output data were analyzed after data collection, and the professional interface was not used or assessed in this study.

Figure 1. The system process of our computerized clinical decision support system (Klinik Access). In this study, "finish form information" was the request to complete the positive version of the System Usability Scale questionnaire, after which the patients were directed to wait for the triage nurse to call them in. Customer PRO UI was not used in this study, and case information was only archived for the study purposes. AI: artificial intelligence; Customer PRO UI: Customer Professional Interface.



Usability

Usability was evaluated by using the positive version of the SUS questionnaire [13]. The questionnaire has 10 items that are rated on a 5-point Likert scale. The difference between the original and positive versions of the questionnaire is that the positive version has only positive statements about usability, whereas in the original one, every other statement is negative [13]. The positive version of the SUS questionnaire was used, as respondents and researchers are less likely to make mistakes when filling in and analyzing the questionnaire, respectively [13]. After completing the web-based CDSS form, the patients filled in the positive version of the SUS questionnaire, and they had the opportunity to give free-text feedback by using the web-based form.

Statistical Analysis

The estimated required sample size for this study (estimated via a power analysis) was a minimum of 246 patients (confidence level=95%; margin of error=5%; study population proportion=80%). A patient was excluded from the analyses if they had intentionally ignored filling in the majority of the web-based form, misused the form, or stated that they were not willing to participate in this study in the free-text field. The

sensitivity and positive predictive values, with 95% CIs, were calculated for the urgency evaluations between the CDSS and the two assessing physicians, and the mean of these results was used to report the performance of the CDSS in comparison to that of the assessing physicians' evaluations. When evaluating the performance of the assessing physicians, the sensitivity was calculated for both assessors, and the mean was used to report the performance of the assessors.

The urgency of the patients' needs was evaluated, using 4 categories, by the CDSS and the assessors. However, as the patients were practically grouped only into 2 groups (nonurgent and urgent) in the emergency department, we dichotomized the urgency as nonurgent (blue=0; green=1) and urgent (yellow=2; red=3) cases for the analyses.

The diagnostic accuracy of the CDSS was evaluated, using the differential diagnosis proposed by the system. A condition had to exceed the fixed probability threshold of 15% to be included in the differential diagnosis list. The diagnostic match was evaluated as "exact" if the actual *ICD-10* diagnosis from electronic health records was included in the differential diagnosis list proposed by the CDSS. The diagnostic match was evaluated as "close" if the condition proposed by the CDSS in the differential diagnosis list was a close match to the actual *ICD-10*–coded diagnosis (Table 2). The statistical software IBM SPSS Statistics version 25 (IBM Corporation) and Microsoft Excel 2022 (Microsoft Corporation) were used.

Table . The prevalence of the most frequent *International Statistical Classification of Diseases, Tenth Revision* diagnoses from the electronic health records of the whole study population and the prevalence of the most frequent diagnoses with exact and close matches to those produced by the clinical decision support system (CDSS).

Diagnosis		Frequency
Diagnoses from electronic health records, n (%)		
	A46 Erysipelas	8 (3.8)
	M54.5 Low back pain	7 (3.3)
	M54.4 Lumbago with sciatica	6 (2.8)
	M79.6 Pain in limb, hand, foot, fingers, and toes	6 (2.8)
	R10.3 Pain localized to other parts of lower abdomen	6 (2.8)
	R10.4 Other and unspecified abdominal pain	6 (2.8)
	K35.9 Acute appendicitis, unspecified	5 (2.3)
	T81.4 Infection following a procedure	5 (2.3)
	R07.4 Chest pain, unspecified	4 (1.9)
	S61.0 Open wound of finger(s) without damage to nail	4 (1.9)
	H16.9 Keratitis, unspecified	3 (1.4)
	I49.9 Cardiac arrhythmia, unspecified	3 (1.4)
	J06.9 Acute upper respiratory infection, unspecified	3 (1.4)
	R06.0 Dyspnea	3 (1.4)
	R10.1 Pain localized to upper abdomen	3 (1.4)
	S52.5 Fracture of lower end of radius	3 (1.4)
	S93.4 Sprain and strain of ankle	3 (1.4)
	T15.0 Foreign body in cornea	3 (1.4)
Diagnoses with exact matches to the CDSS, n		
	R10.3 Pain localized to other parts of lower abdomen	6
	A46 Erysipelas	5
	M54.5 Low back pain	5
	R10.4 Other and unspecified abdominal pain	5
	M79.6 Pain in limb, hand, foot, fingers, and toes	4
	S61.0 Open wound of finger(s) without damage to nail	4
	H16.9 Keratitis, unspecified	3
Diagnoses with close matches to the CDSS, n		
	M54.4 Lumbago with sciatica	3

Results

Study Population Characteristics

There were 259 patients who had completed the CDSS form and had comparable information in the electronic health records. Of these patients, 11 were excluded (4 patients were aged younger than 18 y, 5 patients had an emergency department visit that was solely associated with an earlier emergency department or control visit, and 2 patients had an inadequately filled form). Thus, there were 248 patients in total, with 122 (49.2%) female patients and 126 (50.8%) male patients, and the mean age was 46 (range 18-82) years. The median number of symptoms provided by patients to the CDSS form was 5 (range 1-20).

Urgency

Physician A evaluated all 248 patient cases. However, physician B missed 12 patient cases due to difficulties in accessing appropriate electronic health record data, thus leaving 236 comparable patient cases in total (Tables 3-6). The mean sensitivities were 85% for urgent cases and 19% for nonurgent

cases when assessing the performance of the CDSS evaluations in comparison to those of the physicians. The corresponding mean sensitivities were 85% and 35%, respectively, when comparing the evaluations of the assessing physicians. The CDSS did not miss any cases that were evaluated to be emergencies by physicians, that is, all emergency cases evaluated by the physicians were evaluated as either urgent cases or emergency cases by the CDSS (Tables 3-6).

Table . The urgency evaluations of the computerized decision support system (CDSS) and physician A (cases: n=248).[a]

		Physician A cases, n			
		Self-care	Nonurgent	Urgent	Emergency
CDSS cases, n					
	Self-care	3	4	5	0
	Nonurgent	2	4	23	0
	Urgent	2	32	97	4
	Emergency	0	14	54	4

[a]The positive predictive values for nonurgent and emergency cases were 32% (95% CI 20%-46%) and 77% (95% CI 74%-79%), respectively. The sensitivity values for nonurgent and emergency cases were 21% (95% CI 12%-34%) and 85% (95% CI 79%-90%), respectively.

Table . The urgency evaluations of the computerized decision support system (CDSS) and physician B (cases: n=236).[a]

		Physician B cases, n			
		Self-care	Nonurgent	Urgent	Emergency
CDSS cases, n					
	Self-care	0	4	8	0
	Nonurgent	0	1	25	0
	Urgent	0	17	107	5
	Emergency	0	18	55	6

[a]The positive predictive values for nonurgent and emergency cases were 13% (95% CI 6%-26%) and 87% (95% CI 85%-89%), respectively. The sensitivity values for nonurgent and emergency cases were 17% (95% CI 6%-35%) and 84% (95% CI 78%-89%), respectively.

Table . The urgency evaluations of physician A and physician B (cases: n=236; physician A vs physician B).[a]

		Physician B cases, n			
		Self-care	Nonurgent	Urgent	Emergency
Physician A cases, n					
	Self-care	0	4	3	0
	Nonurgent	0	10	42	0
	Urgent	0	16	147	6
	Emergency	0	0	3	5

[a]The sensitivity values for nonurgent and emergency cases were 47% (95% CI 28%-66%) and 78% (95% CI 72%-84%), respectively.

Table . The urgency evaluations of physician A and physician B (cases: n=236; physician B vs physician A).[a]

		Physician A cases, n			
		Self-care	Nonurgent	Urgent	Emergency
Physician B cases, n					
	Self-care	0	0	0	0
	Nonurgent	4	10	16	0
	Urgent	3	42	147	3
	Emergency	0	0	6	5

[a]The sensitivity values for nonurgent and emergency cases were 24% (95% CI 14%-37%) and 91% (95% CI 86%-95%), respectively.

Differential Diagnosis

Of 248 patients, 35 had to be excluded from the differential diagnosis analyses due to the absence of an *ICD-10* diagnosis in the electronic health records or for having a code denoting a Z-diagnosis (factors influencing health status or contact with health services). Thus, 213 cases in total were included in the differential diagnosis analyses. The results of the differential diagnoses of the CDSS are shown in Table 7. The CDSS had an exact match accuracy of 45.5% (97/213), with an additional close match in 12.7% (27/213) of patients. The most frequent actual diagnoses from the electronic health records, including exact- and close-match diagnoses, are presented in Table 2. Other close-match diagnosis examples are H43.8 (other disorders of the vitreous body; CDSS suggestion: visual disturbances), K64.0 (first-degree hemorrhoids; CDSS suggestion: bleeding from anus), and H16.0 (corneal ulcer; CDSS suggestion: foreign object in the eye). The median number of diagnoses (ie, conditions) within the differentials provided by our CDSS was 5 (range 0-21).

Table . The results of the differential diagnostics of the clinical decision support system (CDSS), including possible explanatory factors for missed diagnoses.

Differential diagnosis		CDSS accuracy (diagnoses: N=213), n (%)
Exact match		97 (45.5)
Close match		27 (12.7)
Missed diagnoses		89 (41.8)
	False location	9 (4.2)
	Inadequate response	25 (11.7)
	No identified diagnosis[a]	16 (7.5)
	Other miss	39 (18.3)

[a]The exact *International Statistical Classification of Diseases, Tenth Revision* diagnosis was not included in the CDSS differential diagnostics selection.

Usability

A positive version of the SUS questionnaire was answered by 95.4% (247/259) of the whole study population. The mean SUS score for the CDSS was 78.2 (SD 16.8). A total of 31 patients had given feedback on the CDSS, of whom 12 gave positive feedback, 14 had some critiques or suggested certain changes, and 5 gave neutral feedback. The most frequent free-text feedback for improving the CDSS addressed possible challenges for older patients (n=5) and the need for guidance to fill in the form (n=3).

Discussion

Overview

To the authors' knowledge, this is the first real-world study that uses a large population with a wide age range to extensively evaluate the performance and usability of a CDSS for urgency evaluation and differential diagnosis in a patient setting. The CDSS was found to be equally as sensitive as the assessing physicians in terms of urgent patient evaluation, but the CDSS underperformed in nonurgent patient cases. The CDSS was considered to be safe, as none of the emergency cases evaluated by physicians were evaluated as nonurgent cases by the CDSS. The diagnostic performance of the CDSS was good, with a 45.5% (97/213) exact match accuracy and 12.7% (27/213) close match accuracy for differential diagnosis in a real-world setting, and it included a vast range of possible diagnoses. The usability of the CDSS was considered to be impressive, with a mean SUS score of 78.2 (SD 16.8).

Digital health technologies and CDSSs have the potential to optimize health care resources and enable patients to manage their conditions themselves more effectively [4]. In addition, CDSS technologies have been suggested to have a significant role in helping with both patient management and triage [17]. Digital health technologies, including CDSSs, are however often lacking an evidence base [18]. Both the importance and the lack of validation for digital health technologies have been acknowledged widely [2,19].

With web-based CDSS access, patients can complete their inquiries at home or prior to their emergency department arrival. Health care and triage personnel would then have comprehensive written information on the patient's symptoms, demographics, and history data available when the patient enters the emergency department. This can help to speed up the work in triage considerably. A CDSS could also categorize patient cases as urgent and nonurgent cases when patients enter the emergency department. This would improve the waiting times for patients with urgent conditions and thus improve treatment outcomes [1].

Even though any CDSS should not be too risk averse [4], it is of utmost importance that the CDSS is safe enough to avoid missing any emergency cases. Considering the high resource demands and massive time pressures within the emergency department, directing even every fifth patient safely to nonurgent care would be very beneficial. This is a significant factor to consider for the patient population who would, in any case, enter the emergency department, regardless of possible overtriaging. Eventually, the purpose of a CDSS in the emergency department is to relieve the health care personnel's burden, so that time and effort can be optimally allocated [20,21]. Additionally, to fully gain the benefits of a CDSS, it is important that the CDSS is properly implemented in health care professionals' everyday work and workflow [22].

Urgency Evaluation

In this study, the CDSS was found to be as sensitive as physicians in identifying urgent patient cases. For nonurgent patient cases, the CDSS was oversensitive. This is logical when considering the nature of artificial intelligence–driven digital health technology in the health care field, where safety is of utmost importance. As described previously, there is a great need to validate every CDSS in real-world clinical settings [4,7].

The original purpose of this study was to assess the triage nurses' assessments, using the Emergency Severity Index (ESI) [23]. However, the distribution of the ESI scores was greatly skewed; no patients had an ESI score of 1, and only 2% (5/248) of patients had an ESI score of 2, which substantially limited data analysis. Therefore, the physicians were considered to be the most suited to assessing the urgency of the patients' conditions by using the triage nurse's information.

There is a lack of studies assessing CDSSs for urgency evaluation in a real-world setting with a broad range of possible conditions [7,8]. Cotte et al [24] assessed the triage performance of a CDSS (Ada) in a real-world emergency department setting with 344 patients. They found overtriage in 57% of cases and undertriage in 9% of cases when the CDSS was compared to physicians' evaluations. Yu et al [25] assessed 2 CDSSs (ie, symptom checkers) retrospectively for 100 real-world patients that entered the emergency department, and they found a sensitivity of 58%; however, this study was performed retrospectively by researchers and thus lacked an acute emergency department setting and patient usability evaluation. Schmieding et al [26] assessed the performance of 22 CDSS technologies in 2015 and 2020 and discovered no improvement in triage performance over the 5-year period. However, to conclude, studies assessing CDSS performance underline the importance of real-world data. This is because clinical vignettes have been found to have considerable inherent limitations when used to assess diagnostic accuracy or triage safety, in comparison to real-world data [5]. Just recently, Fraser et al [8] evaluated the performance and usability of a CDSS (Ada; ie, in a similar setting to ours) in an emergency department. In terms of urgency, among 37 patient cases, 22% and 14% were considered unsafe and too risky by at least one or two out of three physicians, respectively [8].

Differential Diagnosis

When assessing the CDSS's diagnostic performance with *ICD-10* diagnoses set by physicians in the patients' electronic health records, an exact diagnostic match was found in 45.5% (97/213) of cases, and missed diagnoses were found in 41.8% (89/213) of cases. Given the real-world, acute emergency department study setting, this can be considered a good performance. As described previously, studies assessing the performance of a CDSS usually use vignettes without real patient data, which makes it difficult to draw any clinical conclusions [4,5]. In a review by Wallace et al [7], there were only a few studies that assessed diagnostic accuracy by using real patient data, and this was done only in specific subspecialties. Despite these facts, the accuracy of the primary diagnosis was found to be low, with a range of 19% to 38% [7]. In a recent emergency department study that used another CDSS, the sensitivity of the differential diagnosis was far higher (70%) [8].

This study assessed the exact and close diagnostic matches between our CDSS and electronic health record diagnoses. Through this, we wanted to highlight the challenges faced when using real-world data. Even though the CDSS evaluated in this study contained over 500 medical conditions, patients may experience rare conditions that are difficult and often futile to include within CDSS differential diagnostic algorithms. Among 213 patients, there were 16 patients (7.5%) who had an *ICD-10* diagnosis that was not included in the CDSS differential diagnostics, such as myelodysplastic syndrome (D46.9). As described previously, there are numerous diagnoses that may not be necessary to include in CDSS differential diagnostics. This is because some diagnoses may be too niche for emergency department triage or are only relevant for certain specialist clinical settings, such as tertiary subspecialty care.

The main objective of any CDSS is to provide support for triage evaluation, as setting an accurate diagnosis often requires specific tests, such as blood sample and imaging tests [7]. Therefore, it is impractical to aim for the perfect diagnostic performance of a CDSS. Nevertheless, a CDSS could optimally guide health care professionals to consider a patient's relevant differential diagnosis.

Unlike several CDSSs, the one used in this study (Klinik Access) proposed a differential diagnostic list by setting a threshold (15%) for the probabilities of relevant conditions. This mimics the testing threshold design in a physician's clinical decision-making, during which all of the relevant and possible conditions are considered [27,28].

With regard to practical issues, there are multiple factors that affect a unique patient case in real life and are complicated to include in theoretical studies with clinical vignettes. For instance, in this study, some patients were referred to the emergency department by the general practitioner, which affected the answers of the CDSS questionnaire. Further, as in real life, some diagnoses were not accurately recorded within the electronic health records by the physicians in the emergency department, although cases with empty diagnoses and cases with a code denoting a Z-diagnosis were excluded.

Usability

Usability and user interaction have been noted as key elements for evaluating a CDSS [2,29]. In this study, the CDSS (Klinik Access) achieved a mean SUS score of 78.2 (SD 16.8), which indicates good usability and is in the highest quartile when evaluating the SUS [30]. There is a lack of studies that have evaluated the usability of a CDSS in a real-world setting with a diverse patient sample that actually used the CDSS themselves. In addition, our study population of 247 patients is fairly large when compared to other study populations, considering the real-world setting of our study. Fraser et al [8] showed the acceptable usability of a CDSS that used the Technology Acceptance Model among 40 emergency department patients. Additionally, in a study consisting of 49 psychotherapy outpatients, the CDSS (Ada) achieved a mean SUS score of

81.5 [6]. However, this study was performed in a narrow patient context with a more limited study population.

With regard to gathered usability feedback, the fact that the patients with an acute ongoing condition used the CDSS while waiting for a triage nurse in the emergency department is an important detail to consider, and this activity could have negatively impacted the results. Patients filled in the CDSS web-based form at the emergency department waiting room while experiencing an acute ongoing condition and not, for example, at home without distractions, as in the typical primary care setting process. Although participation was voluntary, according to a few answers, some patients had low motivation to fill in the form, most likely as they were aware that the CDSS would not have affected their actual care in any way. Additionally, some answers in the CDSS questionnaire were clearly inadequate. Certain patients had difficulties with choosing the right location for their symptoms and conditions. Some patients also omitted some of their symptoms in the selection phase but reported those symptoms in the free-text field instead, which left the algorithm with insufficient information. These factors are partly inevitable when evaluating patients in an emergency department setting. However, motivation issues could have affected the results due to the study design, in which the CDSS was separate from the actual care. These limitations underline again the challenges of CDSS studies in a real-world setting compared to those of studies using clinical vignettes.

As the positive version of the SUS questionnaire was added solely for the research purposes, it was collected by using a separate web-based form after patients completed the CDSS form, and it did not include the medical information of the patients. Therefore, we could not evaluate the usability among different user categories, such as age groups or gender. Nevertheless, the mean SUS score describes well the results of the whole study population, as 95.4% (247/259) of the study population completed the SUS questionnaire.

There has been no reported statistical difference between the original and positive versions of the SUS [31]. This study was performed in the emergency department; thus, the ease of answering the questionnaire is relevant for the patients [13]. In addition, the positive version of the SUS has been suggested for users with cognitive load or stress, which is typical in an emergency department setting [31].

Strengths and Limitations of This Study

There are clear strengths in this study. We want to underline the fact that, unlike the majority of the previous studies that explored CDSSs and symptom checkers (with data provided by the patients themselves), the population sample in this study consisted of real patients who entered the emergency department of a university hospital, and patient CDSS data were entered via the internet by the patients themselves [4,7]. To the authors' knowledge, this is the first study to assess the diagnostic performance and triage sensitivity of a CDSS in a broad clinical setting, using patient-submitted data. The patient sample is representative of the normal adult patient population entering the emergency department in Finland and is presumably unbiased by demographic variance, which strengthens the findings of this study [20].

This study also has some limitations. Even though this study included a broad set of patients, it excluded children, adolescents, and patients with cognitive impairments. This was due to the fact that these patient groups may need another person to fill in the web-based CDSS form, which would have likely affected the study results. In fact, unlike the majority of CDSS technologies, the CDSS assessed in this study (Klinik Access) does include children and adolescents, and a study is in progress to assess the diagnostic and triage performance of the CDSS for children and adolescents aged <18 years in a university hospital setting. Further, the patient selection could have been biased, as the patients were voluntarily participating in this study. We also excluded patients who arrived to the emergency department via prehospital emergency care (ambulance). Additionally, 49.2% (122/248) of patients were female and 50.8% (126/248) were male in this study, and the mean age was 46 (range 18-82) years. Therefore, the patient population in this study reflects the typical patient sample in a university hospital emergency department. The fact that the evaluations of the assessing physicians were reliant on the recordings of a triage nurse can also be considered as a limitation. One assessing physician experienced challenges in accessing the appropriate electronic health record data for 12 patients, which were due to the access permissions used within the electronic health records of the emergency department.

The patients were informed that this study would not affect their normal assessment or the care of their condition. This could have lowered patients' motivation to complete the web-based CDSS form, which in turn could have resulted in inadequacies in answers. If true, this would directly diminish the study results. These issues are unavoidable when dealing with real-world patient data and, again, underline the challenges of CDSS studies using real-world data compared to those of vignette studies. To tackle these limitations, further studies should assess health care professionals' practical use of a CDSS to show the performance and possible benefits of the CDSS.

Conclusions

This study is the first to validate a CDSS with a large, real-world patient population within a university hospital emergency department. The CDSS was found to be safe to use, with no missed urgent cases, and was equally as sensitive as emergency physicians' judgments in urgent patient cases. It provided acceptable differential diagnoses and good patient usability, as evaluated via a positive version of the SUS questionnaire. The CDSS should further be evaluated for its practical use (eg, studies in which health care professionals can use the CDSS and its output in real time). This would likely further demonstrate the practical benefits and effectiveness of CDSS technologies in emergency medicine.

Acknowledgments

The authors thank Maarit Kumpulainen (Kuopio University Hospital Living Lab) for contributing to the design and practical arrangements of this study. This work was supported by Klinik Healthcare Solutions Oy (provided tablets and salary to research assistants who recruited the study population).

Conflicts of Interest

JM, RL, NH, and PH are employed by Klinik Healthcare Solutions Oy, the creator of the clinical decision support system (CDSS). JM, RL, and PH have stock ownership in Klinik Healthcare Solutions Oy. The independent emergency care physicians who evaluated patients alongside our CDSS were not employed by Klinik Healthcare Solutions Oy.

References

1. Fernandes M, Vieira SM, Leite F, Palos C, Finkelstein S, Sousa JMC. Clinical decision support systems for triage in the emergency department using intelligent systems: a review. Artif Intell Med 2020 Jan;102:101762. [doi: 10.1016/j.artmed.2019.101762] [Medline: 31980099]
2. Fraser H, Coiera E, Wong D. Safety of patient-facing digital symptom checkers. Lancet 2018 Nov 24;392(10161):2263-2264. [doi: 10.1016/S0140-6736(18)32819-8] [Medline: 30413281]
3. Chan F, Lai S, Pieterman M, Richardson L, Singh A, Peters J, et al. Performance of a new symptom checker in patient triage: Canadian cohort study. PLoS One 2021 Dec 1;16(12):e0260696. [doi: 10.1371/journal.pone.0260696] [Medline: 34852016]
4. Gottliebsen K, Petersson G. Limited evidence of benefits of patient operated intelligent primary care triage tools: findings of a literature review. BMJ Health Care Inform 2020 May;27(1):e100114. [doi: 10.1136/bmjhci-2019-100114] [Medline: 32385041]
5. El-Osta A, Webber I, Alaa A, Bagkeris E, Mian S, Sharabiani MTA, et al. What is the suitability of clinical vignettes in benchmarking the performance of online symptom checkers? An audit study. BMJ Open 2022 Apr 27;12(4):e053566. [doi: 10.1136/bmjopen-2021-053566] [Medline: 35477872]
6. Hennemann S, Kuhn S, Witthöft M, Jungmann SM. Diagnostic performance of an app-based symptom checker in mental disorders: comparative study in psychotherapy outpatients. JMIR Ment Health 2022 Jan 31;9(1):e32832. [doi: 10.2196/32832] [Medline: 35099395]
7. Wallace W, Chan C, Chidambaram S, Hanna L, Iqbal FM, Acharya A, et al. The diagnostic and triage accuracy of digital and online symptom checker tools: a systematic review. NPJ Digit Med 2022 Aug 17;5(1):118. [doi: 10.1038/s41746-022-00667-w] [Medline: 35977992]
8. Fraser HSF, Cohan G, Koehler C, Anderson J, Lawrence A, Pateña J, et al. Evaluation of diagnostic and triage accuracy and usability of a symptom checker in an emergency department: observational study. JMIR Mhealth Uhealth 2022 Sep 19;10(9):e38364. [doi: 10.2196/38364] [Medline: 36121688]
9. Singh H, Meyer AND, Thomas EJ. The frequency of diagnostic errors in outpatient care: estimations from three large observational studies involving US adult populations. BMJ Qual Saf 2014 Sep;23(9):727-731. [doi: 10.1136/bmjqs-2013-002627] [Medline: 24742777]
10. Singh H, Giardina TD, Meyer AND, Forjuoh SN, Reis MD, Thomas EJ. Types and origins of diagnostic errors in primary care settings. JAMA Intern Med 2013 Mar 25;173(6):418-425. [doi: 10.1001/jamainternmed.2013.2777] [Medline: 23440149]
11. Shackel B. Usability – context, framework, definition, design and evaluation. Interact Comput 2009 Dec;21(5-6):339-346. [doi: 10.1016/j.intcom.2009.04.007]
12. Brooke J. SUS: A 'quick and dirty' usability scale. In: Jordan PW, Thomas B, Weerdmeester BA, McClelland IL, editors. Usability Evaluation in Industry. London, United Kingdom: Taylor & Francis; 1996. ISBN: 9780429157011.
13. Sauro J, Lewis JR. When designing usability questionnaires, does it hurt to be positive? 2011 Presented at: CHI '11: CHI Conference on Human Factors in Computing Systems; May 7-11; Vancouver, BC. [doi: 10.1145/1978942.1979266]
14. Maramba I, Chatterjee A, Newman C. Methods of usability testing in the development of eHealth applications: a scoping review. Int J Med Inform 2019 Jun;126:95-104. [doi: 10.1016/j.ijmedinf.2019.03.018] [Medline: 31029270]
15. Spiegelhalter DJ, Myles JP, Jones DR, Abrams KR. Bayesian methods in health technology assessment: a review. Health Technol Assess 2000;4(38):1-130. [Medline: 11134920]
16. Akobeng AK. Understanding diagnostic tests 1: sensitivity, specificity and predictive values. Acta Paediatr 2007 Mar;96(3):338-341. [doi: 10.1111/j.1651-2227.2006.00180.x] [Medline: 17407452]
17. Elias P, Damle A, Casale M, Branson K, Churi C, Komatireddy R, et al. A web-based tool for patient triage in emergency department settings: validation using the emergency severity index. JMIR Med Inform 2015 Jun;10(2):e23. [doi: 10.2196/medinform.3508] [Medline: 26063343]
18. Greaves F, Joshi I, Campbell M, Roberts S, Patel N, Powell J. What is an appropriate level of evidence for a digital health intervention? Lancet 2019 Dec 22;392(10165):2665-2667. [doi: 10.1016/S0140-6736(18)33129-5] [Medline: 30545779]
19. The Lancet. Is digital medicine different? Lancet 2018 Jul 14;392(10142):95. [doi: 10.1016/S0140-6736(18)31562-9] [Medline: 30017135]

20. Israni ST, Verghese A. Humanizing artificial intelligence. JAMA 2019 Jan 1;321(1):29-30. [doi: 10.1001/jama.2018.19398] [Medline: 30535297]
21. Lin SY, Mahoney MR, Sinsky CA. Ten ways artificial intelligence will transform primary care. J Gen Intern Med 2019 Aug;34(8):1626-1630. [doi: 10.1007/s11606-019-05035-1] [Medline: 31090027]
22. Kujala S, Hörhammer I. Health care professionals' experiences of web-based symptom checkers for triage: cross-sectional survey study. J Med Internet Res 2022 May 5;24(5):e33505. [doi: 10.2196/33505] [Medline: 35511254]
23. Wuerz RC, Milne LW, Eitel DR, Travers D, Gilboy N. Reliability and validity of a new five-level triage instrument. Acad Emerg Med 2000 Mar;7(3):236-242. [doi: 10.1111/j.1553-2712.2000.tb01066.x] [Medline: 10730830]
24. Cotte F, Mueller T, Gilbert S, Blümke B, Multmeier J, Hirsch MC, et al. Safety of triage self-assessment using a symptom assessment app for walk-in patients in the emergency care setting: observational prospective cross-sectional study. JMIR Mhealth Uhealth 2022 Mar 28;10(3):e32340. [doi: 10.2196/32340] [Medline: 35343909]
25. Yu SWY, Ma A, Tsang VHM, Chung LSW, Leung SC, Leung LP. Triage accuracy of online symptom checkers for accident and emergency department patients. Hong Kong Journal of Emergency Medicine 2020 Jul;27(4):217-222. [doi: 10.1177/1024907919842486]
26. Schmieding ML, Kopka M, Schmidt K, Schulz-Niethammer S, Balzer F, Feufel MA. Triage accuracy of symptom checker apps: 5-year follow-up evaluation. J Med Internet Res 2022 May 10;24(5):e31810. [doi: 10.2196/31810] [Medline: 35536633]
27. Cahan A, Gilon D, Manor O, Paltiel O. Probabilistic reasoning and clinical decision-making: do doctors overestimate diagnostic probabilities? QJM 2003 Oct;96(10):763-769. [doi: 10.1093/qjmed/hcg122] [Medline: 14500863]
28. Pauker SG, Kassirer JP. The threshold approach to clinical decision making. N Engl J Med 1980 May 15;302(20):1109-1117. [doi: 10.1056/NEJM198005153022003] [Medline: 7366635]
29. World Health Organization. Monitoring and Evaluating Digital Health Interventions: A Practical Guide to Conducting Research and Assessment. Geneva, Switzerland: World Health Organization; 2016. ISBN: 9789241511766.
30. Bangor A, Kortum P, Miller J. Determining what individual SUS scores mean: adding an adjective rating scale. J Usability Stud 2009 May 1;4(3):114-123. [doi: 10.5555/2835587.2835589]
31. Kortum P, Acemyan CZ, Oswald FL. Is it time to go positive? Assessing the positively worded System Usability Scale (SUS). Hum Factors 2021 Sep;63(6):987-998. [doi: 10.1177/0018720819881556] [Medline: 31913715]

Abbreviations

CDSS: clinical decision support system
ESI: Emergency Severity Index
***ICD-10*:** *International Statistical Classification of Diseases, Tenth Revision*
SUS: System Usability Scale

Edited by C Perrin, C Lovis; submitted 24.02.23; peer-reviewed by D Zhao, P Gupta, Z Galavi; revised version received 22.06.23; accepted 14.07.23; published 31.08.23.

Please cite as:
Määttä J, Lindell R, Hayward N, Martikainen S, Honkanen K, Inkala M, Hirvonen P, Martikainen TJ
Diagnostic Performance, Triage Safety, and Usability of a Clinical Decision Support System Within a University Hospital Emergency Department: Algorithm Performance and Usability Study
JMIR Med Inform 2023;11:e46760
URL: https://medinform.jmir.org/2023/1/e46760
doi:10.2196/46760

An Ontology-Based Approach to Improving Medication Appropriateness in Older Patients: Algorithm Development and Validation Study

Elena Calvo-Cidoncha[1], PharmD; Julián Verdinelli[2], PharmD; Javier González-Bueno[3], PharmD, PhD; Alfonso López-Soto[4], MD, PhD; Concepción Camacho Hernando[1], MBA, PharmD; Xavier Pastor-Duran[2], MD, PhD; Carles Codina-Jané[1], PharmD, PhD; Raimundo Lozano-Rubí[2], BSCS, MD, PhD

1

2

3

4

Corresponding Author:
Elena Calvo-Cidoncha, PharmD

Abstract

Background: Inappropriate medication in older patients with multimorbidity results in a greater risk of adverse drug events. Clinical decision support systems (CDSSs) are intended to improve medication appropriateness. One approach to improving CDSSs is to use ontologies instead of relational databases. Previously, we developed OntoPharma—an ontology-based CDSS for reducing medication prescribing errors.

Objective: The primary aim was to model a domain for improving medication appropriateness in older patients (chronic patient domain). The secondary aim was to implement the version of OntoPharma containing the chronic patient domain in a hospital setting.

Methods: A 4-step process was proposed. The first step was defining the domain scope. The chronic patient domain focused on improving medication appropriateness in older patients. A group of experts selected the following three use cases: medication regimen complexity, anticholinergic and sedative drug burden, and the presence of triggers for identifying possible adverse events. The second step was domain model representation. The implementation was conducted by medical informatics specialists and clinical pharmacists using Protégé-OWL (Stanford Center for Biomedical Informatics Research). The third step was OntoPharma-driven alert module adaptation. We reused the existing framework based on SPARQL to query ontologies. The fourth step was implementing the version of OntoPharma containing the chronic patient domain in a hospital setting. Alerts generated from July to September 2022 were analyzed.

Results: We proposed 6 new classes and 5 new properties, introducing the necessary changes in the ontologies previously created. An alert is shown if the Medication Regimen Complexity Index is ≥40, if the Drug Burden Index is ≥1, or if there is a trigger based on an abnormal laboratory value. A total of 364 alerts were generated for 107 patients; 154 (42.3%) alerts were accepted.

Conclusions: We proposed an ontology-based approach to provide support for improving medication appropriateness in older patients with multimorbidity in a scalable, sustainable, and reusable way. The chronic patient domain was built based on our previous research, reusing the existing framework. OntoPharma has been implemented in clinical practice and generates alerts, considering the following use cases: medication regimen complexity, anticholinergic and sedative drug burden, and the presence of triggers for identifying possible adverse events.

(JMIR Med Inform 2023;11:e45850) doi:10.2196/45850

KEYWORDS

biological ontologies; decision support systems; inappropriate prescribing; elderly; medication regimen complexity; anticholinergic drug burden; trigger tool; clinical; ontologies; pharmacy; medication; decision support; pharmaceutic; pharmacology; chronic condition; chronic disease; domain; adverse event; ontology-based; alert

Introduction

Medical advances have resulted in a rise of life expectancy. The prevalence of multimorbidity, which is defined as the coexistence of 2 or more chronic conditions, tends to be higher among older people [1]. As a result, the use of multiple medicines, which is commonly referred to as *polypharmacy*, has become a common phenomenon in this population [2,3].

Polypharmacy increases the risk of inappropriate medication [4,5], leading to a greater risk of adverse drug events (ADEs) [6]. ADEs are associated with hospital admissions, higher mortality rates, and increased health care expenditures [7-10]; therefore, improving medication appropriateness in older patients with multimorbidity is a priority [11].

One approach to improving medication appropriateness is to use clinical decision support systems (CDSSs) for assistance during the prescription process. CDSSs are intended to improve health care delivery by enhancing medical decisions with targeted clinical knowledge and patient information [12,13]. Relational databases are the predominant choice when it comes to designing a CDSS. However, due to the main challenges of CDSSs, such as the lack of interoperability or alert fatigue [14-16], there is increasing interest in using ontology-based CDSSs to overcome these challenges. An ontology is an explicit conceptualization of the entities of a domain [17,18]. Because ontologies add semantics to models, they enhance the reusability of data and are more efficient in dealing with changing requirements and maintenance requirements [19-21].

Previously, we used Protégé-OWL (Stanford Center for Biomedical Informatics Research) to develop OntoPharma—an ontology-based CDSS for reducing medication prescribing errors [22]. The domains addressed by OntoPharma include the identification and technical data of medicinal products, as well as data on drug appropriateness for ensuring the safe use of medicines. These domains were addressed in the following four use cases: maximum dosage alerts, a drug-drug interaction checker, renal failure adjustment, and a drug allergy checker. OntoPharma is currently implemented in a tertiary referral hospital.

Alerts generated by OntoPharma are, nowadays, commonly available. To leverage the ease of using ontologies to represent rich and complex knowledge, the modeling of drug knowledge that is absent in usual commercial databases is needed.

For this reason and on the basis of our previous research, the primary aim of this study was to model a domain for improving medication appropriateness in older patients with multimorbidity (hereinafter called the *chronic patient domain*). The secondary aim was to implement the version of OntoPharma containing the chronic patient domain in a hospital setting.

Methods

This study was conducted between 2020 and 2022 at a 710-bed tertiary hospital in Spain, which was equipped with computerized physician order entry (CPOE) and an electronic health record (EHR) system provided by SAP SE. The following 4-step development process was designed: defining the domain scope, representing the domain in the model, adapting the OntoPharma-driven alert module, and implementing the version of OntoPharma containing the chronic patient domain in a hospital setting.

Ethics Approval

This study was approved by the ethics committee of the Hospital Clínic of Barcelona (reference number HCB/2019/0735).

Defining the Domain Scope

Overview of the Domain Scope

The chronic patient domain focused on improving medication appropriateness in older patients with multimorbidity. Given the dimension of the domain, we decided to establish an expert advisory panel to limit the scope of the domain. The group of experts included geriatricians and clinical pharmacists, of whom all were members of the C3RG (Central Catalonia Chronicity Research Group) and had expertise in ensuring medication appropriateness in older patients with multimorbidity. Focus group sessions yielded consensus on the importance of the following three use cases: medication regimen complexity, anticholinergic and sedative drug burden, and the presence of triggers for identifying possible adverse events.

Medication Regimen Complexity

Complex medication regimens are challenging for patients, which may impact medication adherence and safety [23,24]. The Medication Regimen Complexity Index (MRCI), which was developed by George et al [25], is currently the most widely used scale for assessing medication regimen complexity. Medication complexity considers more factors than a simple medication count. The MRCI consists of 65 items, including weighted scores for types of prescribed dosage forms (section A), dosing frequency (section B), and additional administration instructions (section C). The sum of the scores of the three sections provides a total score, with higher scores indicating greater regimen complexity.

The MRCI has been translated and validated for other languages, including Spanish (Spanish MRCI [MRCI-E]) [26]. We used the MRCI-E as a source of information.

Section A provides weights for 32 dosage form and administration route combinations. For example, an oral tablet medication is given a weight of 1. More complex combinations result in higher weights.

Section B provides weights for 23 dosing frequencies (“scheduled” or “as needed”). The “once daily” frequency is used as the baseline (weight of 1), on which the other weightings are built.

Section C provides weights for 10 additional instructions that a patient may need to follow in adhering to a prescribed regimen. Additional administration instructions are related to taking medication at specific times, taking medication in relation to food, taking multiple units at one time, and needing to break or crush a tablet or needing to taper or increase a dose.

Anticholinergic and Sedative Drug Burden

Anticholinergic burden is defined as the cumulative effect of taking 1 or more drugs that are capable of causing anticholinergic adverse effects, and the load increases with the number of medications prescribed [27]. Anticholinergic toxicity is a common problem in older people. Anticholinergic effects are associated with peripheral manifestations (urinary retention, constipation, decreased secretions, etc) and central manifestations (delirium, cognitive disorders, and functional disorders) [27,28].

Several tools have been developed to estimate anticholinergic burden by giving a score to drugs according to their anticholinergic potential [29]. The Drug Burden Index (DBI) is the only scale that accounts for a patient's dose [30]. In addition, the DBI considers not only anticholinergic effects but also sedative effects. The total DBI exposure is calculated as the sum of exposure to any DBI medication, according to the following formula:

(1)

where "D" is the daily dose taken and "δ" is the minimum effective daily dose for that drug.

Byrne et al [31] provided a master DBI list containing a final list of DBI medications and their minimum effective daily doses. The master DBI list included 156 entries. Each entry consisted of the following fields: drug description (ingredient), World Health Organization Anatomical Therapeutic Classification codes, anticholinergic and sedative effects, and minimum effective daily dose (expressed as mg) by route of administration (parenteral, sublingual, buccal, transdermal, rectal, and inhalation).

Triggers

A trigger is defined as a flag, occurrence, or prompt that alerts reviewers to initiate further in-depth investigations regarding a patient's record to determine the presence or absence of an adverse event [32]. An example of a trigger is a potassium level of <2.9 mEq/L in a patient with loop diuretics. Triggers are based on the assumption that any new condition may be due to the use of a drug. Multiple sets of triggers have been developed. Guzmán et al [33] identified the most appropriate triggers for detecting ADEs in older patients with multiple chronic conditions.

The trigger set developed by Guzmán et al [33] included a total of 51 entries. Each entry consisted of the following fields: high-alert medications for patients with chronic illnesses (therapeutic class or ingredient) and triggers for detecting potential ADEs (11 care module triggers, 9 antidote- and treatment-based triggers, 11 medication concentration–based triggers, 18 triggers based on abnormal laboratory values, and 1 emergency department trigger).

Domain Model Representation

Data sets were not organized in a predefined format. Prior to modeling the chronic patient domain through ontologies, we processed all of the information in a relational database to clean the data, detect redundancies, and detect relationships between different concepts.

To add this new domain to OntoPharma, we built on our previous research by using the existing framework, which was composed of 3 ontologies (*Drugs*, *Decision support system* [*DSS*], and *Local pharmacy*) [22]. The *Drugs* ontology was designed to represent the identification and technical data of medicinal products. The *DSS* ontology provides data on drug appropriateness. The *Local pharmacy* ontology was designed to represent local concepts from EHRs and CPOE in order to ensure interoperability.

The design, development, and maintenance of the chronic patient domain was driven by medical informatics specialists and clinical pharmacists. The information was represented in the Web Ontology Language (OWL) [34]. For encoding the OWL ontologies, we used the Protégé 3.5 editor tool [35]. The concepts of the chronic patient domain were organized hierarchically, following a top-down approach, as we did with all previous domains of OntoPharma. The development of the class hierarchy, the defining of properties, and the slotting of concepts were carried out at the same time. Finally, we defined individual instances of the classes represented.

OntoPharma-Driven Alert Module Adaptation

We reused the OntoPharma-driven alert module that was proposed in our previous research [22]. The integration between the CPOE system and the ontologies was performed through a REST API. A REST API call was published (in JSON format) each time a clinician added a new medication in the CPOE system, modified an existing one, or requested on-demand CDSS information. The request contained patient-specific clinical data. SPARQL (Apache Jena Fuseki server) was used to query ontologies [36]. After applying the queries, a returning REST API, with the results, was published.

It was necessary to update the content of the REST API published each time OntoPharma was triggered. We specifically had to add more laboratory parameters (to date, the only one considered was glomerular filtration rate). New local concepts were manually mapped with existing concepts in the ontologies. In addition, we created new SPARQL queries to ensure the safe use of medicines in older patients.

Alerts were shown in the CPOE system in cases of high medication regimen complexity, in cases of high anticholinergic and sedative drug burden, or in cases where triggers for detecting ADEs in older patients were present. In addition, patients were required to be older than 65 years.

In accordance with the recommendations of end users, the user interface proposed in the previous paper [22] was slightly modified to ensure usability and minimal interference with the clinician's workflow.

Formal testing was performed to demonstrate that the new version of the ontology-driven alert module met functional requirements. In addition, clinical pharmacists performed manual testing in a control environment (the SAP quality assurance server) to evaluate whether the alert module functioned properly when generating the prescribing alerts.

Implementation of the Version of OntoPharma Containing the Chronic Patient Domain in a Hospital Setting

In July 2022, the version of OntoPharma containing the chronic patient domain was implemented at one ward of the internal medicine unit, which had capacity for 20 admissions. Informatics staff and clinical pharmacists were responsible for the diffusion and for providing support.

A retrospective analysis of the alerts generated was performed. We included patients who were admitted to the internal medicine ward from July to September 2022. The following patient data were collected: gender, age, duration of hospital stay, and number of medications during hospital stay. We further examined the alerts, including the number of alerts, the types of alerts, clinical relevance, and the acceptance rates.

Quantitative variables were expressed as means and SDs for variables with a normal distribution or as medians and IQRs for variables with a skewed distribution. Qualitative variables were expressed as percentages. Data analysis was carried out by using SPSS 20.0 (IBM Corp).

Results

Knowledge Representation Using Ontologies

Overview of Ontologies

For modeling the chronic patient domain, we proposed new classes and properties, introducing the necessary changes in the ontologies previously created (*Drugs*, *DSS*, and *Local pharmacy*). The three ontologies are interconnected. The import schema of ontologies is shown in Figure 1.

Figures 2-4 provide diagrams showing the relationships between classes for defining the chronic patient domain in the *Drugs* ontology, *DSS* ontology, and *Local pharmacy* ontology, respectively.

Multimedia Appendix 1 contains a list of the medication knowledge concepts and their definitions, which were used to define the chronic patient domain. Multimedia Appendix 2 contains a list of properties and their facets, which were also used to define the chronic patient domain.

Figure 1. Import schema of the ontologies used in OntoPharma. For modeling drug-related knowledge, 3 ontologies have been developed (*Drugs*, *DSS*, and *Local pharmacy*). Each ontology has been divided into 2 parts. The first part provides concepts and classes (also known as *T-Box*), and the second provides the instances of these concepts (also known as *A-Box*). The three ontologies are interconnected, as shown in Figure 1. DSS: Decision support system.

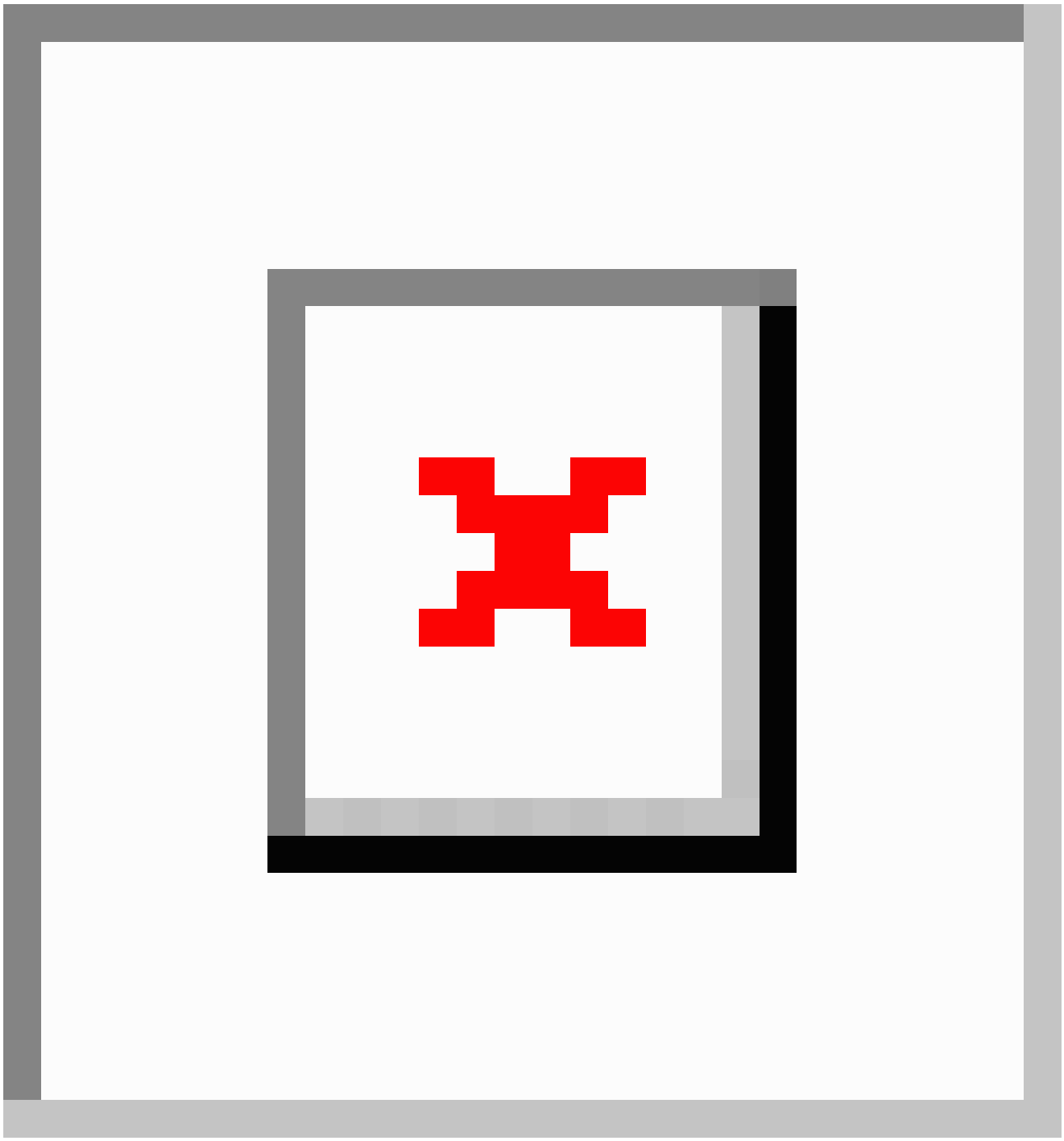


Figure 2. Diagram showing the relationships between classes in the *Drugs* ontology for defining the chronic patient domain. The *Drugs* ontology was designed to represent the identification and technical data of medicinal products.

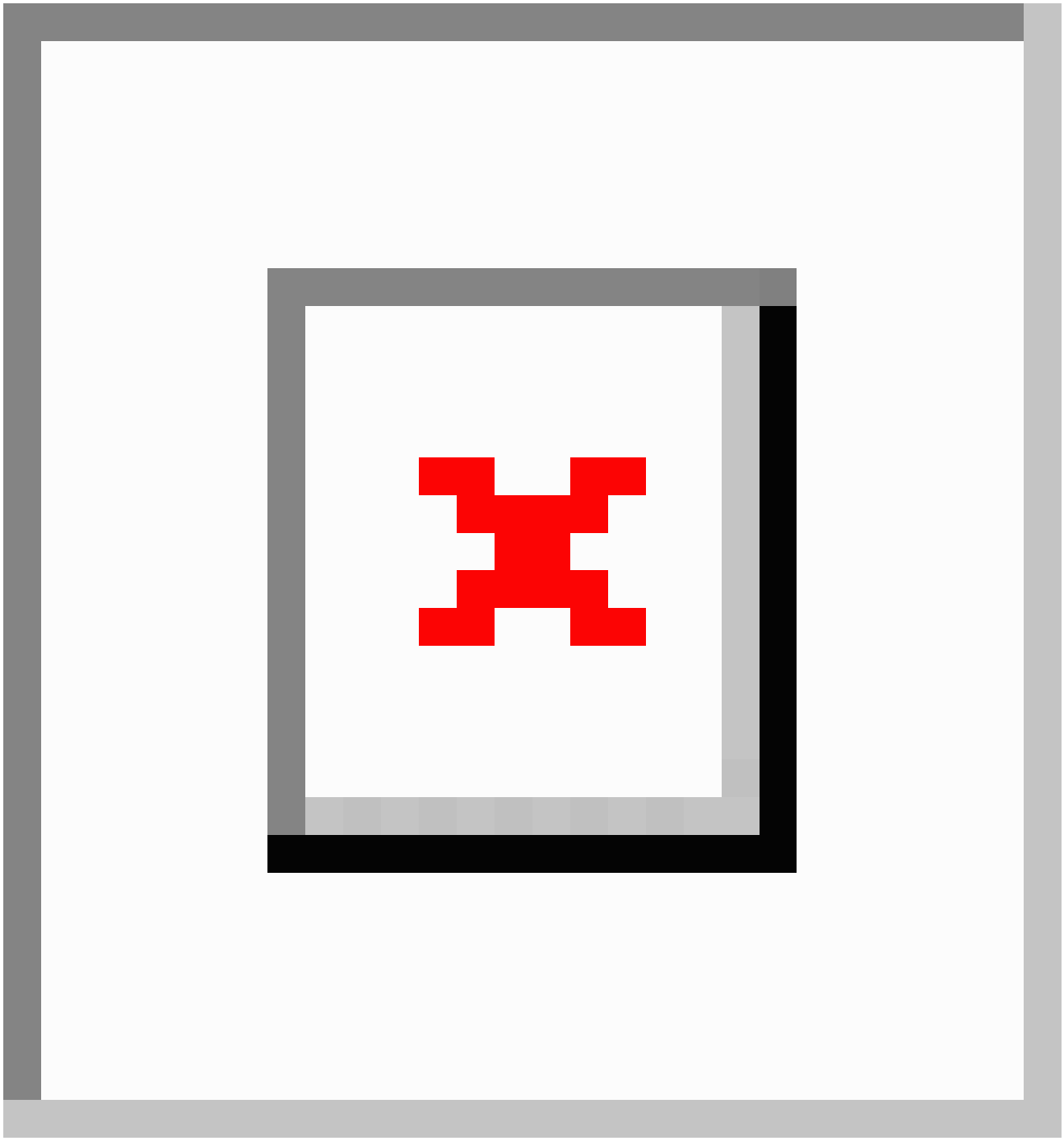


Figure 3. Diagram showing the relationships between classes in the *DSS* ontology for defining the chronic patient domain. The DSS ontology provides data on drug appropriateness. Concepts that were specifically created to define the chronic patient domain are highlighted in yellow. DBI: Drug Burden Index; DSS: Decision support system; MRCI: Medication Regimen Complexity Index.

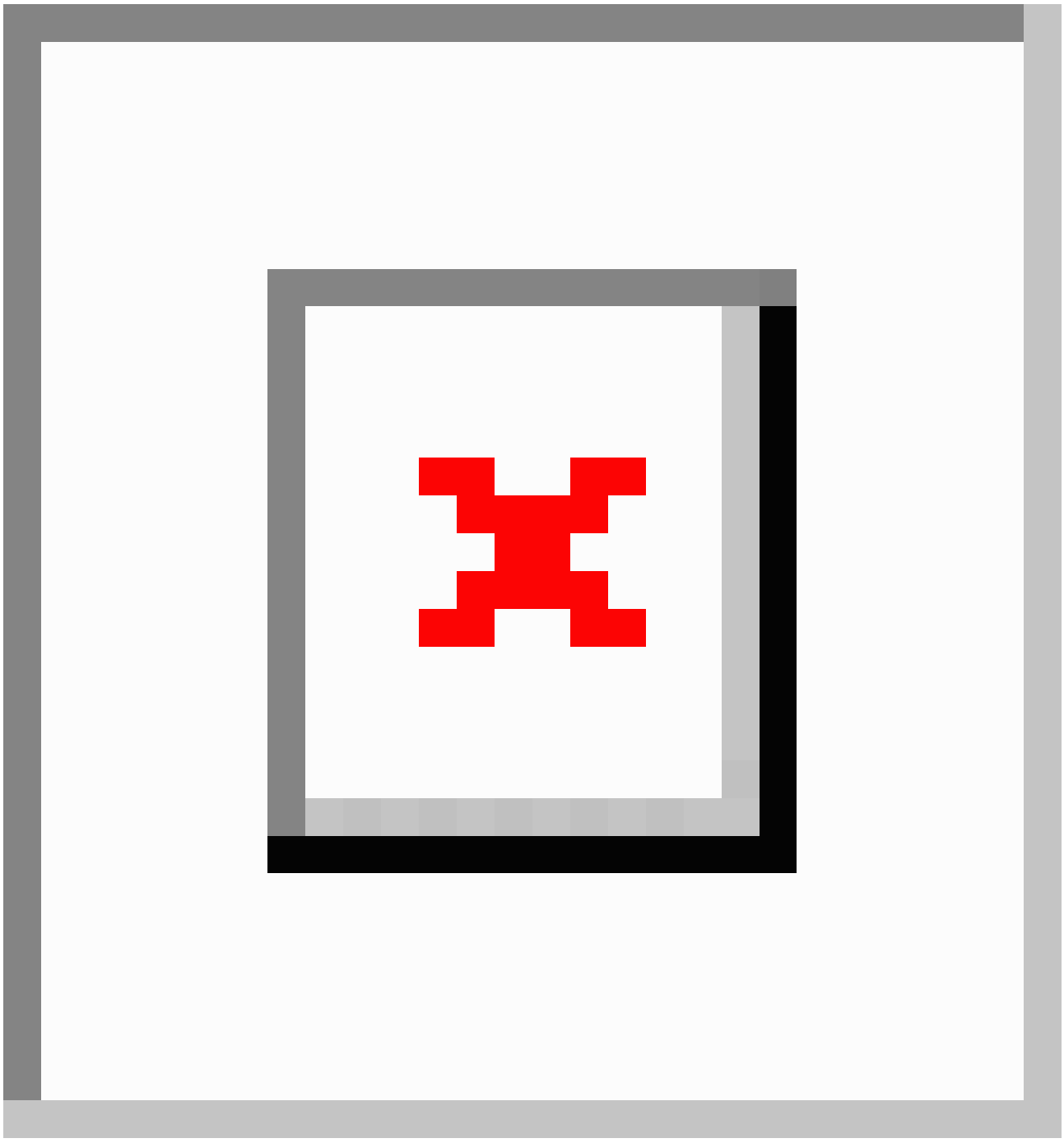
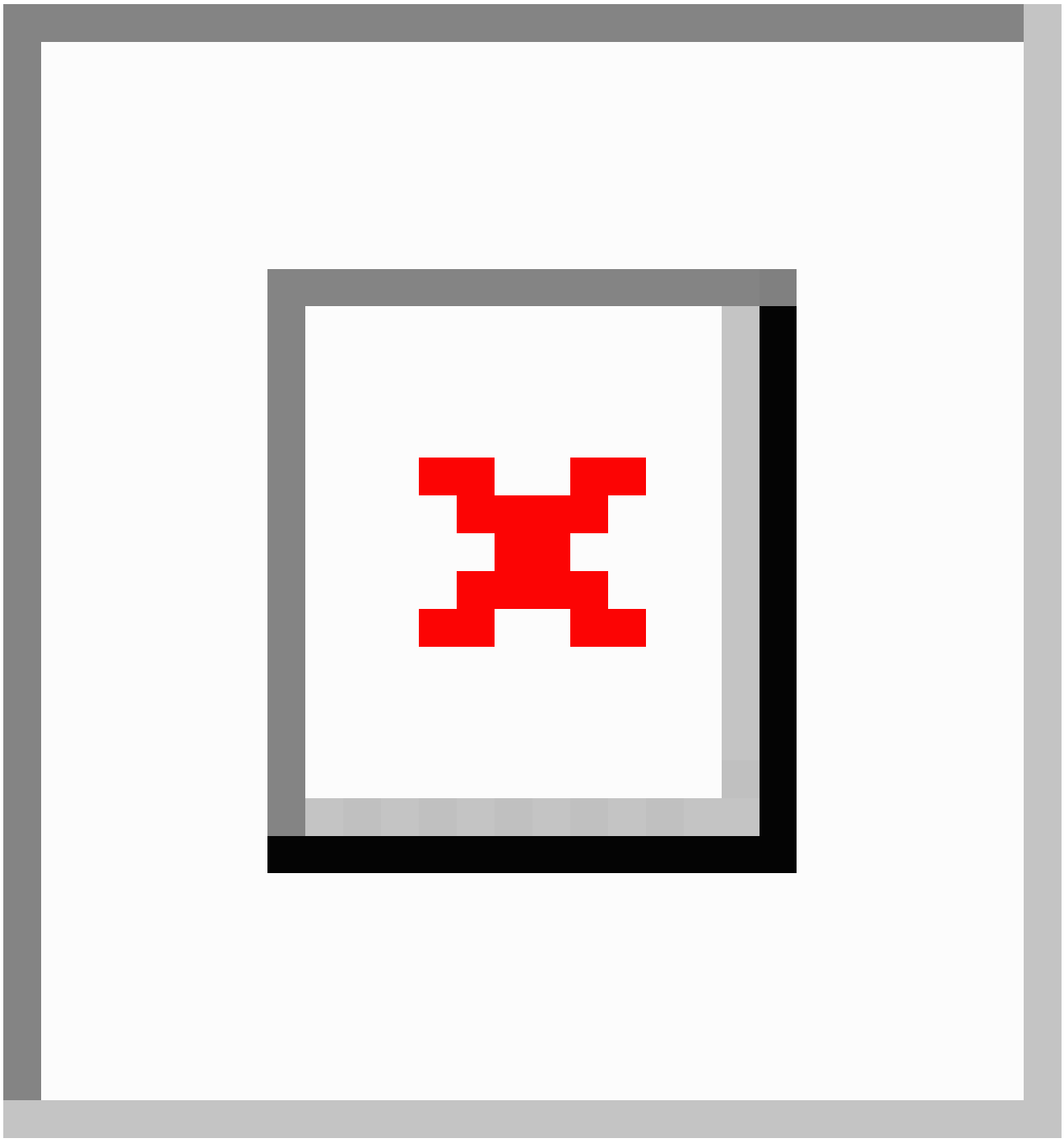


Figure 4. Diagram showing the relationships between classes in the *Local pharmacy* ontology for defining the chronic patient domain. The *Local pharmacy* ontology was designed to represent local concepts from electronic health records and computerized physician order entry. Each local concept is mapped to the corresponding OntoPharma concept.



Medication Regimen Complexity

The MRCI quantifies drug regimen complexity based on dosage form, dosage frequency, and additional instructions.

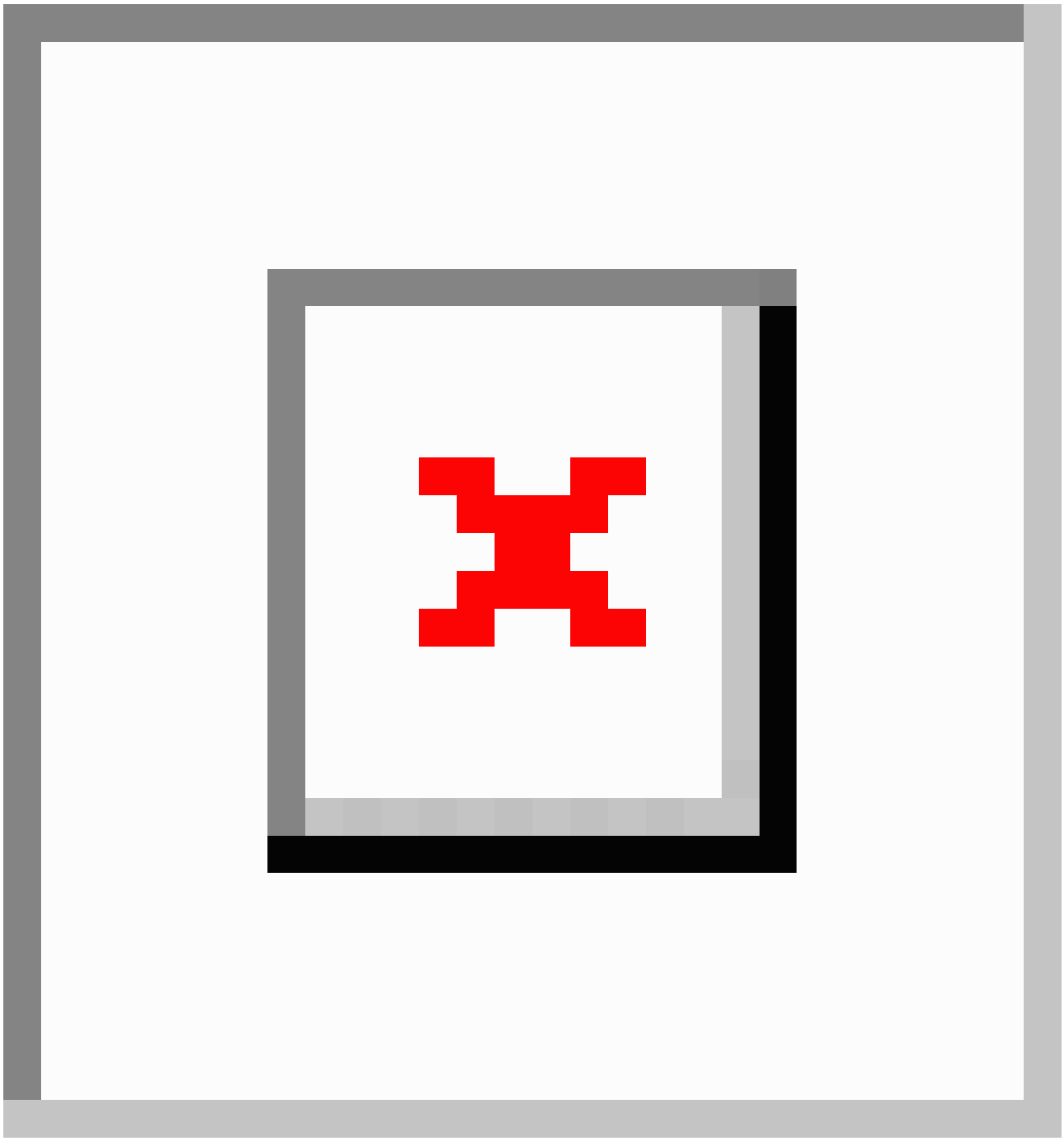
To represent the weighted scores for types of prescribed dosage forms (section A), we first created the concept "MRCI A form" (*DSS* ontology), which comprises 30 subclasses for identifying the possible dosage form and route of administration combinations. To provide the weight for each dosage form and route of administration combination, we introduced the property "mrci A weight."

To represent the weighted scores for dosage frequency (section B), we introduced the following two attached properties within the class "Local frequency" (*Local pharmacy* ontology): "mrci B," which provides weights for "scheduled" dosing frequencies, and "mrci B PRN," which provides weights for "as needed" dosing frequencies.

We identified 231 distinct frequency combinations. Frequency weights were assigned, considering that frequency data also contained indicators that qualify for component C scoring, such as indicators to take medication less often than once per day (eg, once every 48 hours) or indicators to take medication at specific times (before a meal, at bedtime, etc).

To represent the weighted scores for additional administration instructions (section C) related to taking medication with or without food, we introduced 1 attached property ("mrci C") within the class "Virtual medicinal product (VMP)" (*Drugs* ontology). A virtual medicinal product is an abstract representation of an active medicinal ingredient associated with strength information and a route of administration (eg, "omeprazole 20mg capsule"). We assigned a total of 6257 weights. Figure 5 provides a class diagram to model medication regimen complexity, which is explained with an example.

Figure 5. Class diagram to model medication regimen complexity. Circles represent the classes needed to quantify drug regimen complexity. Squares represent the attached properties (object or data) within each class. An example is given in brackets. MRCI: Medication Regimen Complexity Index; VMP: virtual medicinal product.



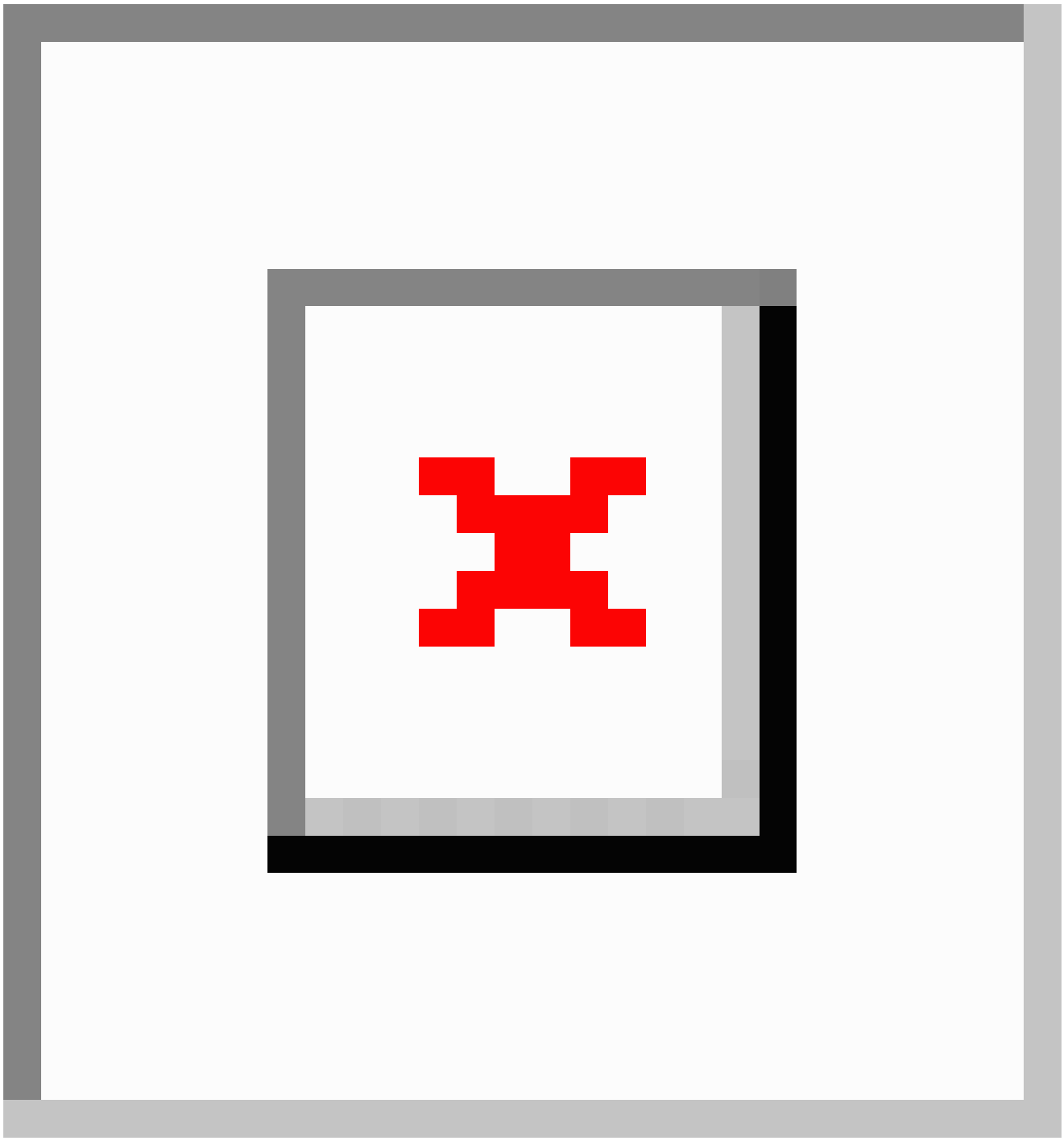
Anticholinergic and Sedative Drug Burden

The OWL concept that was used to enter the data on anticholinergic and sedative drug burden was "DBI" (*DSS* ontology), in reference to the scale used for its calculation. Because the DBI is a dose‐related measure of anticholinergic and sedative drug exposure, we created the "DBI" concept as a subclass of the "Dose appropriateness" concept. To provide enough information to calculate the DBI, we introduced the property "medd," which describes the minimum effective daily dose of each drug.

The "DBI" class contains 164 individuals. Each individual contains the following knowledge: ingredient (eg, alprazolam),

route of administration (eg, oral), age range (eg, 65-999 years), minimum effective daily dose (eg, 0.5), unit (eg, mg), base unit (eg, every 24 hours), and alert-related data. Figure 6 provides a class diagram to model anticholinergic and sedative drug burden, which is explained with an example.

Figure 6. Class diagram to model anticholinergic and sedative drug burden. Circles represent the classes needed to quantify anticholinergic and sedative drug burden. Squares represent the attached properties (object or data) within each class. An example is given in brackets. DBI: Drug Burden Index.

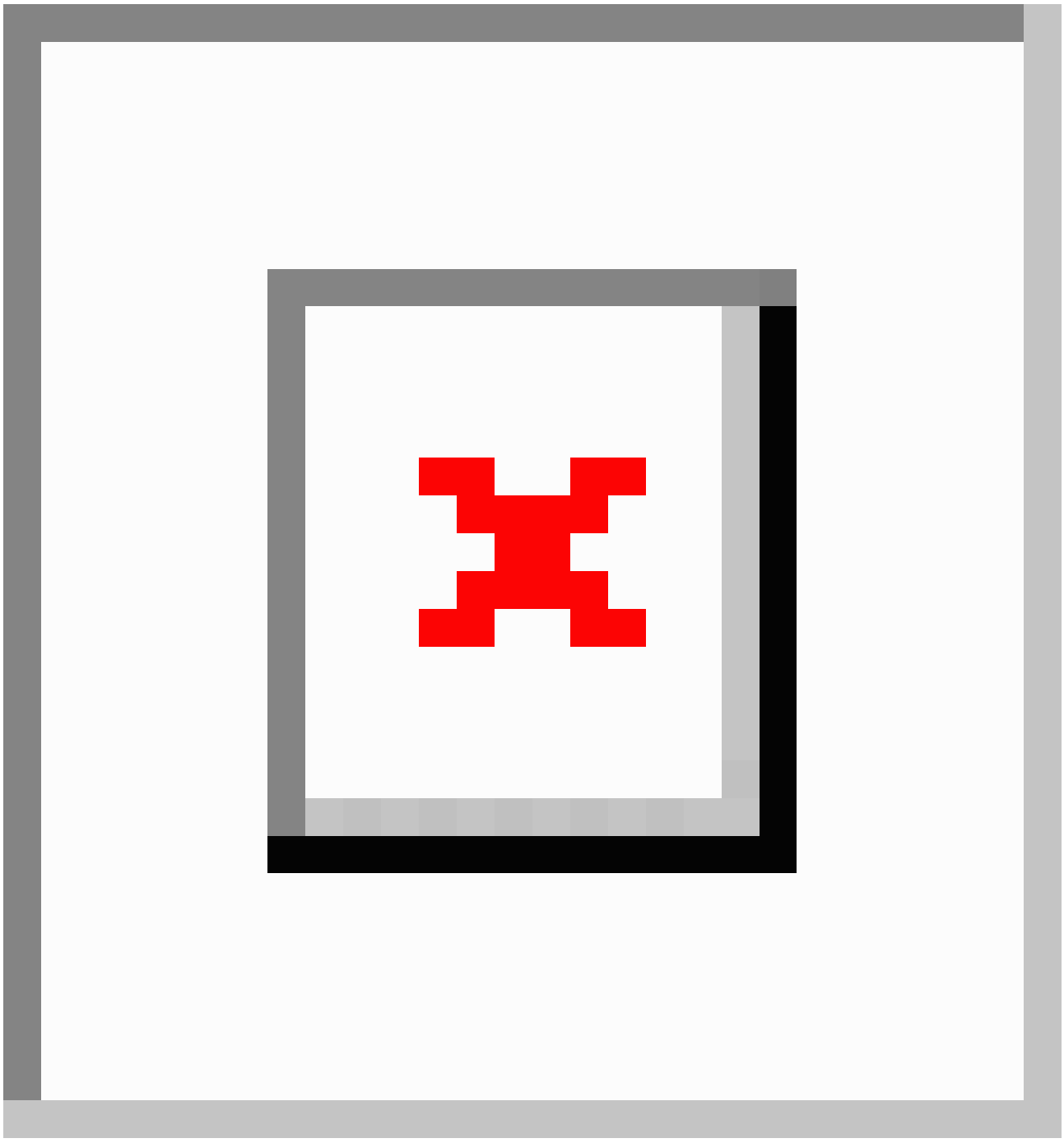


Triggers

The OWL concept that was used to enter the triggers for detecting ADEs in older patients with multiple chronic conditions was "Trigger tool" (*DSS* ontology). We created the "Trigger tool" concept as a subclass of the "Appropriateness lab test" concept because we only included triggers based on abnormal laboratory values. Introducing new properties was not required.

The "Trigger tool" class contains 821 individuals. Each individual contains the following knowledge: ingredient (eg, furosemide), route of administration (eg, parenteral), age range (eg, 65-999 years), lab test (eg, serum glucose), lab test unit (eg, mg/dL), low value (eg, 0), high value (eg, 110), and alert-related data. Figure 7 provides a class diagram to model triggers for detecting ADEs in patients with multimorbidity, which are explained with an example.

Figure 7. Class diagram to model triggers for detecting adverse drug events in patients with multimorbidity. Circles represent the classes needed to identify triggers of a possible adverse drug event. Squares represent the attached properties (object or data) within each class. An example is given in brackets.



Alerts

In addition to the abovementioned actions, we made changes related to the "Alert" class (*DSS* ontology). The "Alert" class included the information that was displayed when appropriateness criteria were not met. The following fields related to the "Alert" class remained unchanged: alert description, alert recommendation, alert source, alert date (the date when the alert was last updated), and related information (supporting documentation). In addition to the existing instances of the "Alert level" class ("not recommended," "contraindicated," or "unallowed prescription"), we created a new one called "risk minimization." We also introduced a new class—the "Drug intervention" subclass of the "Alert" class (*DSS* ontology). We defined the following two types of drug intervention: "global drug" intervention for when a complete treatment revision is required (eg, high MRCI) and "specific drug" intervention for when a partial treatment revision is required (eg, high DBI).

Knowledge Not Represented Using Ontologies

With regard to medication regimen complexity, we did not represent complexity based on the additional instructions related to taking multiple units at one time or needing to break or crush a tablet.

The triggers used to identify ADEs can be abnormal laboratory values, the use of certain medications or antidotes, or changes in clinical status that may indicate a possible medication-related harm. We only represented triggers based on abnormal laboratory values.

OntoPharma-Driven Alert Module Adaptation

The OntoPharma-driven alert module works the same as it did in our previous study [22]. Once the patient-specific clinical data are sent from the CPOE system and EHR to the ontologies, local concepts are matched to their equivalent OntoPharma concepts. With regard to the chronic patient domain, we defined the following decision rules.

The MRCI is obtained by summing the scores of the three sections.

The section A score is estimated by considering the dosage form and route of administration combination. Because administering the same dosage form more than once is easier than administering different dosage forms, each dosage form and route of administration combination is counted only once within a regimen. For example, if a patient's regimen consists of taking 3 tablets orally, their component A score is 1, not 3.

Section B and C scores are estimated by considering the dosage frequency and the virtual medicinal product prescribed, respectively. The cutoff point selected for triggering an alert (MRCI≥40) was determined by the expert advisory panel, in accordance with the literature.

The total DBI is calculated with the equation  , where "D" is the daily dose taken by the individual patient and "δ" is the minimum effective daily dose for that drug. The daily dose taken for each DBI medication is estimated by considering the dose, dose unit, and frequency. We have defined conversion factors for cases where the drug dose unit prescribed is different from the unit dose defined in the ontologies. The minimum effective daily dose is represented in the *DSS* ontology for each ingredient and route of administration combination. The cutoff point selected for triggering an alert (DBI≥1) was determined by the advisory panel, in accordance with the literature.

To evaluate the presence of triggers for identifying ADEs, we consider the ingredient regardless of the dosage. If a patient has several laboratory values, we consider the most recent values. An alert is triggered when a value is outside of the defined range [33].

Medications prescribed "as needed" were not considered in previous cases.

With regard to the interface, alerts are shown in different colors (red, orange, and yellow) according to their clinical relevance (contraindicated, moderate relevance, and low relevance). We added a new label (blue) to identify alerts aimed at risk minimization. To date, the advisory text contains the generic drug name, a short description of the possible concern, and a recommendation for improving medication appropriateness. The generic drug name is still displayed if the alert requires "specific drug" intervention. In cases where the alert requires "global drug" intervention, the text "Review total treatment" is displayed. We also included a hyperlink to relevant literature.

Alerts related to the chronic patient domain were defined, such as soft-stop alerts, so that the clinician can decide whether to ignore or accept the alert. To avoid alert fatigue, if an alert is ignored once, it will not be displayed again.

The interface that displays the alerts also includes a link to a user guide and an activity registry that serves as traceability system.

Despite the addition of new use cases, the results show that the response time for generating decision support remains short (within milliseconds), with minimal impact on the user's workflow.

Implementation of the Version of OntoPharma Containing the Chronic Patient Domain in a Hospital Setting

A total of 107 patients were included. The median age was 86 (IQR 80-90) years, and the majority of patients were women (n=63, 58.9%). The median length of hospital stay was 8 (IQR 5-13) days. Patients had a median of 15 (IQR 11-19) medications.

Of the 107 patients, 96 (89.7%) received at least one alert. OntoPharma generated 364 alerts (mean 3.9, SD 5.3 alerts per patient). Of these, 296 (81.3%) alerts were considered of low relevance, and 68 (18.7%) aimed at risk minimization. Further, 154 (42.3%) alerts were accepted.

Details of the types of alerts and the acceptance rates are included in Table 1. The most frequent alerts were alerts due to high anticholinergic and sedative drug burden (231/364, 63.5%), followed by alerts due to high medication regimen complexity (68/364, 18.7%) and alerts due to the presence of triggers (65/364, 17.8%).

Table . Description of the types of alerts generated by OntoPharma and the acceptance rates.

Type of alert	Frequency (N=364), n (%)	Acceptance rate, n (%[a])
Medication regimen complexity	68 (18.7)	40 (58.8)
Anticholinergic and sedative drug burden	231 (63.5)	84 (36.4)
Triggers	65 (17.8)	30 (46.2)

[a]Percentages were calculated by using the numbers in the "Frequency" column as denominators.

Discussion

Principal Results

This paper presents a modeling approach, which was formalized in ontological terms, for defining the chronic patient domain that provides support for improving medication appropriateness in older patients with multimorbidity. The chronic patient domain was built on OntoPharma—an ontology-based CDSS for reducing medication prescribing errors that has already been implemented in a tertiary referral hospital [22].

There are already ontology-based CDSSs that address medication management in patients with chronic conditions [37]. However, to the best of our knowledge, this is the first ontology-based approach that models medication regimen complexity, anticholinergic and sedative drug burden, and triggers for identifying possible adverse events. Farrish and Grando [38] built an ontology to assist with the management of polypharmacy prescriptions for patients with multiple chronic conditions to reduce the overall treatment complexity. Recently, Román-Villarán et al [39] developed an ontology-based CDSS for patients with complex chronic conditions. However, the knowledge sources used were different from ours, including clinical practice guidelines, the LESS-CHRON (List of Evidence-Based Deprescribing for Chronic Patients) criteria, and the STOPP/START (Screening Tool of Older Persons' Prescriptions and Screening Tool to Alert to Right Treatment) criteria, among others. These ontology approaches for patients with chronic conditions have been validated with patient data from databases. However, it is important to note that they are not implemented in a real environment, unlike our ontology approach. OntoPharma provides rapid and real-time support to improve medication appropriateness in older patients with multimorbidity.

Using ontologies instead of relational databases, which are the predominant choice in current commercial CDSSs, has distinct advantages [40,41]. First, the semantic approach and the use of OWL enable a convenient infrastructure for reuse. Hence, we reused the existing OntoPharma framework, without having to start from scratch. In addition, ontologies are more flexible and efficient in dealing with changes; thus, it was possible to add a new domain to OntoPharma without major complications. We were able to model a complex domain, creating only 6 new classes and 5 new properties. This was possible because the three ontologies (*Drugs*, *DSS*, and *Local pharmacy*) are interconnected (Figure 1), and classes are linked between them through object properties.

To ensure flexibility, scalability, and sustainability, we operated on the most appropriate level of abstraction. To define anticholinergic drug burden and triggers, we considered the ingredient. However, to define weighted scores for additional administration instructions (MRCI section C), we considered the class "Virtual medicinal product (VMP)."

To integrate structured clinical data with clinical knowledge, we reused the mappings previously established in the ontology *Local pharmacy*. It was only necessary to add some new mappings related to laboratory parameters.

End users participated throughout the development of the chronic patient domain in order to ensure usability and gain user acceptance [42,43]. As a result, we introduced some changes in the user interface, such as new clinical relevance levels and a hyperlink to relevant literature. Some proposals for improvement, such as showing the laboratory values next to the alert, have not been implemented yet. Usability may also be influenced by the response time for generating decision support. However, response time has not been modified, showing that OntoPharma is scalable.

Limitations

In terms of evaluation, we have not identified a database-based system for direct comparison with OntoPharma. van der Sijs et al [44] conducted a systematic review, concluding that drug safety alerts are overridden by clinicians in 49% to 96% of cases. Our acceptance rate (154/364, 42.3%) was expected to be better, considering that an expert advisory panel selected the most useful information to improve medication appropriateness in older patients. This may be partly explained by the following limitations. First, appropriateness criteria were evaluated if patients were older than 65 years. Considering that the older population is heterogeneous, we should also have considered frailty—a known factor indicative of vulnerability to medication-related problems [45]. Second, the alerts with a lower acceptance rate (84/231, 36.4%) were related to the DBI. Interventions for reducing the DBI commonly involve progressive medication deprescribing, which is difficult to realize in a tertiary hospital and would be easier in intermediate care [46]. On the other hand, poor adherence is one of the major consequences of high MRCI scores [23]. In hospitals, the administration of medications is primarily the nurses' responsibility; therefore, clinicians may have not given sufficient importance to MRCI alerts. Acceptance rates might improve in outpatient care. Our research focused primarily on clinician decision-making. The variables analyzed allowed us to identify the scale of potentially inappropriate medications and the usefulness of OntoPharma. However, evaluating OntoPharma's influence on health outcomes is a challenge that we should take up in future.

As mentioned in our previous paper on OntoPharma [22], one limitation of this study is maintaining the evidence and keeping it relevant and up to date [47]. To create individual instances, we extracted the information from papers. Since this was not done via automatic extraction, it was a time-consuming process. With regard to the maintenance, we must assign a complexity weight if there is a new dose form or dosing frequency; these data are not updated frequently. In addition, we must check if new medications have additional administration instructions, are capable of causing anticholinergic adverse effects, or are included in the set of triggers for detecting ADEs in older patients.

Of note, although mapping in this study did not take excessive time, we are aware that manual mapping is a resource-intensive and ongoing process.

We have not represented all of the knowledge from the sources of information. OntoPharma relies on structured data; therefore, we have prioritized representing data that are structured in text

format within the EHR. As a result, medication regimen complexity may be underestimated because special instructions are underrepresented. In addition, there are triggers that are different from abnormal laboratory values that are not represented in ontologies. Even though there exist large amounts of health care data, the main challenge to improving results of CDSSs is converting free-text data into structured fields computationally [48].

In future implementations, we will continue to represent complex drug knowledge that is absent in commercial databases. We are currently modeling knowledge for supporting the neonatal population and populations at risk for hepatitis B virus reactivation. To capture clinicians' reasoning processes, we must place a high priority on increasing structured patient data within EHRs.

Other areas for future work are mentioned in our previous OntoPharma paper [22], such as developing a more complex CDSS that can be applied across the entire treatment process and is not only restricted to the medication prescription process. Finally, we must continue working on customized alerts to avoid alert fatigue [49].

Despite the limitations, we believe that our methods have been successful in modeling knowledge related to the chronic patient domain and that the proposed version of OntoPharma is an enhancement of the previous one. Although optimizing care in older patients is a context-dependent, complex process, we believe that developing an ontology to support the chronic patient domain constitutes a major step toward improving medication appropriateness in a generalizable and reusable way.

Conclusions

Polypharmacy in the older population poses challenges to the delivery of medical care because of the increased difficulties in guaranteeing appropriate prescription. We proposed an ontology-based approach to provide support for improving medication appropriateness in older patients with multimorbidity in a scalable, sustainable, and reusable way. OntoPharma has been implemented in clinical practice and generates alerts, considering the following use cases: medication regimen complexity, anticholinergic and sedative drug burden, and the presence of triggers for identifying possible adverse events.

Acknowledgments

We thank Central Catalonia Chronicity Research Group (C3RG)-Line of research "Person-Centred Prescription" for providing advice in the conceptualization of the use cases. We especially thank Joan Espaulella-Panicot, Daniel Sevilla-Sánchez, and Núria Molist-Brunet.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Medication knowledge concepts that are represented in OntoPharma to define the chronic patient domain.
[DOCX File, 22 KB - medinform_v11i1e45850_app1.docx]

Multimedia Appendix 2
Properties and their facets, which are represented in OntoPharma to define the chronic patient domain.
[DOCX File, 20 KB - medinform_v11i1e45850_app2.docx]

References

1. Kojima T, Mizokami F, Akishita M. Geriatric management of older patients with multimorbidity. Geriatr Gerontol Int 2020 Dec;20(12):1105-1111. [doi: 10.1111/ggi.14065] [Medline: 33084212]
2. Guthrie B, Makubate B, Hernandez-Santiago V, Dreischulte T. The rising tide of polypharmacy and drug-drug interactions: population database analysis 1995-2010. BMC Med 2015 Apr 7;13:74. [doi: 10.1186/s12916-015-0322-7] [Medline: 25889849]
3. Charlesworth CJ, Smit E, Lee DSH, Alramadhan F, Odden MC. Polypharmacy among adults aged 65 years and older in the United States: 1988-2010. J Gerontol A Biol Sci Med Sci 2015 Aug;70(8):989-995. [doi: 10.1093/gerona/glv013] [Medline: 25733718]
4. Maher RL, Hanlon J, Hajjar ER. Clinical consequences of polypharmacy in elderly. Expert Opin Drug Saf 2014 Jan;13(1):57-65. [doi: 10.1517/14740338.2013.827660] [Medline: 24073682]
5. Sehgal V, Bajwa SJS, Sehgal R, Bajaj A, Khaira U, Kresse V. Polypharmacy and potentially inappropriate medication use as the precipitating factor in readmissions to the hospital. J Family Med Prim Care 2013 Apr;2(2):194-199. [doi: 10.4103/2249-4863.117423] [Medline: 24479078]
6. Hedna K, Hakkarainen KM, Gyllensten H, Jönsson AK, Petzold M, Hägg S. Potentially inappropriate prescribing and adverse drug reactions in the elderly: a population-based study. Eur J Clin Pharmacol 2015 Dec;71(12):1525-1533. [doi: 10.1007/s00228-015-1950-8] [Medline: 26407684]

7. Pedrós C, Formiga F, Corbella X, Arnau JM. Adverse drug reactions leading to urgent hospital admission in an elderly population: prevalence and main features. Eur J Clin Pharmacol 2016 Feb;72(2):219-226. [doi: 10.1007/s00228-015-1974-0] [Medline: 26546335]
8. Marcum ZA, Pugh MJV, Amuan ME, Aspinall SL, Handler SM, Ruby CM, et al. Prevalence of potentially preventable unplanned hospitalizations caused by therapeutic failures and adverse drug withdrawal events among older veterans. J Gerontol A Biol Sci Med Sci 2012 Aug;67(8):867-874. [doi: 10.1093/gerona/gls001] [Medline: 22389461]
9. Conforti A, Costantini D, Zanetti F, Moretti U, Grezzana M, Leone R. Adverse drug reactions in older patients: an Italian observational prospective hospital study. Drug Healthc Patient Saf 2012;4(1):75-80. [doi: 10.2147/DHPS.S29287] [Medline: 22888275]
10. Bates DW, Spell N, Cullen DJ, Burdick E, Laird N, Petersen LA, et al. The costs of adverse drug events in hospitalized patients. Adverse Drug Events Prevention Study Group. JAMA 1997 Jan 22;277(4):307-311. [doi: 10.1001/jama.1997.03540280045032] [Medline: 9002493]
11. González-Bueno J, Espaulella-Panicot J. Tailored care in frail patients with multimorbidity: future prospects. Farm Hosp 2021 Sep 2;45(5):221-222. [Medline: 34806579]]
12. Kuperman GJ, Bobb A, Payne TH, Avery AJ, Gandhi TK, Burns G, et al. Medication-related clinical decision support in computerized provider order entry systems: a review. J Am Med Inform Assoc 2007;14(1):29-40. [doi: 10.1197/jamia.M2170] [Medline: 17068355]
13. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med 2020 Feb 6;3(1):17. [doi: 10.1038/s41746-020-0221-y] [Medline: 32047862]
14. Khairat S, Marc D, Crosby W, Al Sanousi A. Reasons for physicians not adopting clinical decision support systems: critical analysis. JMIR Med Inform 2018 Apr 18;6(2):e24. [doi: 10.2196/medinform.8912] [Medline: 29669706]
15. Ancker JS, Edwards A, Nosal S, Hauser D, Mauer E, Kaushal R, et al. Effects of workload, work complexity, and repeated alerts on alert fatigue in a clinical decision support system. BMC Med Inform Decis Mak 2017 Oct;17(1):36. [doi: 10.1186/s12911-017-0430-8] [Medline: 28395667]
16. Wright A, McEvoy DS, Aaron S, McCoy AB, Amato MG, Kim H, et al. Structured override reasons for drug-drug interaction alerts in electronic health records. J Am Med Inform Assoc 2019 Oct 1;26(10):934-942. [doi: 10.1093/jamia/ocz033] [Medline: 31329891]
17. Gruber TR. A translation approach to portable ontology specifications. Knowl Acquis 1993 Jun;5(2):199-220. [doi: 10.1006/knac.1993.1008]
18. Uschold M, Gruninger M. Ontologies: principles, methods and applications. Knowl Eng Rev 1996 Jun;11(2):93-136. [doi: 10.1017/S0269888900007797]
19. Bodenreider O, Stevens R. Bio-ontologies: current trends and future directions. Brief Bioinform 2006 Sep;7(3):256-274. [doi: 10.1093/bib/bbl027] [Medline: 16899495]
20. Haendel MA, Chute CG, Robinson PN. Classification, ontology, and precision medicine. N Engl J Med 2018 Nov;379(15):1452-1462. [doi: 10.1056/NEJMra1615014] [Medline: 30304648]
21. Musen MA. Scalable software architectures for decision support. Methods Inf Med 1999 Dec;38(4-5):229-238. [doi: 10.1055/s-0038-1634422] [Medline: 10805007]
22. Calvo-Cidoncha E, Camacho-Hernando C, Feu F, Pastor-Duran X, Codina-Jané C, Lozano-Rubí R. OntoPharma: ontology based clinical decision support system to reduce medication prescribing errors. BMC Med Inform Decis Mak 2022 Oct;22(1):238. [doi: 10.1186/s12911-022-01979-3] [Medline: 36088328]
23. Ingersoll KS, Cohen J. The impact of medication regimen factors on adherence to chronic treatment: a review of literature. J Behav Med 2008 Jun;31(3):213-224. [doi: 10.1007/s10865-007-9147-y] [Medline: 18202907]
24. Mansur N, Weiss A, Beloosesky Y. Looking beyond polypharmacy: quantification of medication regimen complexity in the elderly. Am J Geriatr Pharmacother 2012 Aug;10(4):223-229. [doi: 10.1016/j.amjopharm.2012.06.002] [Medline: 22749668]
25. George J, Phun YT, Bailey MJ, Kong DCM, Stewart K. Development validation of the medication regimen complexity index. Ann Pharmacother 2004 Sep;38(9):1369-1376. [doi: 10.1345/aph.1D479] [Medline: 15266038]
26. de la Fuente JS, Diaz AS, Cañamares-Orbis I, Ramila E, Izquierdo-Garcia E, Esteban C, et al. Cross-cultural adaptation and validation of the medication regimen complexity index adapted to Spanish. Ann Pharmacother 2016 Nov;50(11):918-925. [doi: 10.1177/1060028016656385] [Medline: 27371950]
27. Tune LE. Anticholinergic effects of medication in elderly patients. J Clin Psychiatry 2001;62 Suppl 21:11-14. [Medline: 11584981]]
28. Fox C, Richardson K, Maidment ID, Savva GM, Matthews FE, Smithard D, et al. Anticholinergic medication use and cognitive impairment in the older population: the medical research Council cognitive function and ageing study. J Am Geriatr Soc 2011 Aug;59(8):1477-1483. [doi: 10.1111/j.1532-5415.2011.03491.x] [Medline: 21707557]
29. Villalba-Moreno AM, Alfaro-Lara ER, Pérez-Guerrero MC, Nieto-Martín MD, Santos-Ramos B. Systematic review on the use of anticholinergic scales in poly pathological patients. Arch Gerontol Geriatr 2016;62:1-8. [doi: 10.1016/j.archger.2015.10.002] [Medline: 26518612]

30. Hilmer SN, Mager DE, Simonsick EM, Cao Y, Ling SM, Windham BG, et al. A drug burden index to define the functional burden of medications in older people. Arch Intern Med 2007 Apr 23;167(8):781-787. [doi: 10.1001/archinte.167.8.781] [Medline: 17452540]
31. Byrne CJ, Walsh C, Cahir C, Ryan C, Williams DJ, Bennett K. Anticholinergic and sedative drug burden in community-dwelling older people: a national database study. BMJ Open 2018 Jul 6;8(7):e022500. [doi: 10.1136/bmjopen-2018-022500] [Medline: 29982221]
32. Griffin FA, Resar RK. Institute for Healthcare improvement. IHI global trigger tool for measuring adverse events (second edition). 2009. URL: www.ihi.org/resources/Pages/IHIWhitePapers/IHIGlobalTriggerToolWhitePaper.aspx [accessed 2023-01-17]
33. Guzmán MDT, Banqueri MG, Otero MJ, Lara ERA, Lagranja PC, Ramos BS. Development of a trigger tool to identify adverse drug events in elderly patients with multimorbidity. J Patient Saf 2021 Sep 1;17(6):e475-e482. [doi: 10.1097/PTS.0000000000000389] [Medline: 28617720]
34. Deborah L, van Harmelen F. W3C. OWL Web Ontology Language overview. 2004. URL: www.w3.org/TR/owl-features/ [accessed 2023-01-17]
35. Knublauch H, Fergerson RW, Noy NF, Musen MA. The Protégé OWL Plugin: an open development environment for semantic web applications. 2004 Presented at: Presented at Third International Semantic Web Conference; November 7–11; Hiroshima, Japan p. 229-243. [doi: 10.1007/b102467]
36. Sheeba T, Krishnan R. Semantic retrieval based on SPARQL and SWRL for learner profile. Int J Appl Eng Res 2015;10(14):34549-34554.
37. Dissanayake PI, Colicchio TK, Cimino JJ. Using clinical reasoning ontologies to make smarter clinical decision support systems: a systematic review and data synthesis. J Am Med Inform Assoc 2020 Jan 1;27(1):159-174. [doi: 10.1093/jamia/ocz169] [Medline: 31592534]
38. Farrish S, Grando A. Ontological approach to reduce complexity in polypharmacy. AMIA Annu Symp Proc 2013 Nov 16;2013:398-407. [Medline: 24551346]]
39. Román-Villarán E, Alvarez-Romero C, Martínez-García A, Escobar-Rodríguez GA, García-Lozano MJ, Barón-Franco B, et al. A personalized ontology-based decision support system for complex chronic patients: retrospective observational study. JMIR Form Res 2022 Aug 2;6(8):e27990. [doi: 10.2196/27990] [Medline: 35916719]
40. Martinez-Cruz C, Blanco IJ, Vila MA. Ontologies versus relational databases: are they so different? A comparison. Artif Intell Rev 2012 Dec;38(4):271-290. [doi: 10.1007/s10462-011-9251-9]
41. Uschold M. Ontology and database schema: what's the difference. Appl Ontol 2015;10(3-4):243-258. [doi: 10.3233/AO-150158]
42. Baysari MT, Zheng WY, Van Dort B, Reid-Anderson H, Gronski M, Kenny E. A late attempt to involve end users in the design of medication-related alerts: survey study. J Med Internet Res 2020 Mar 13;22(3):e14855. [doi: 10.2196/14855] [Medline: 32167479]
43. Miller K, Mosby D, Capan M, Kowalski R, Ratwani R, Noaiseh Y, et al. Interface, information, interaction: a narrative review of design and functional requirements for clinical decision support. J Am Med Inform Assoc 2018 May 1;25(5):585-592. [doi: 10.1093/jamia/ocx118] [Medline: 29126196]
44. van der Sijs H, Aarts J, Vulto A, Berg M. Overriding of drug safety alerts in computerized physician order entry. J Am Med Inform Assoc 2006;13(2):138-147. [doi: 10.1197/jamia.M1809] [Medline: 16357358]
45. Nguyen QD, Wu C, Odden MC, Kim DH. Multimorbidity patterns, frailty, and survival in community-dwelling older adults. J Gerontol A Biol Sci Med Sci 2019 Jul 12;74(8):1265-1270. [doi: 10.1093/gerona/gly205] [Medline: 30169580]
46. González-Bueno J, Sevilla-Sánchez D, Puigoriol-Juvanteny E, Molist-Brunet N, Codina-Jané C, Espaulella-Panicot J. Improving medication adherence and effective prescribing through a patient-centered prescription model in patients with multimorbidity. Eur J Clin Pharmacol 2022 Jan;78(1):127-137. [doi: 10.1007/s00228-021-03207-9] [Medline: 34448906]
47. Sim I, Gorman P, Greenes RA, Haynes RB, Kaplan B, Lehmann H, et al. Clinical decision support systems for the practice of evidence-based medicine. J Am Med Inform Assoc 2001;8(6):527-534. [doi: 10.1136/jamia.2001.0080527] [Medline: 11687560]
48. Burger G, Abu-Hanna A, de Keizer N, Cornet R. Natural language processing in pathology: a scoping review. J Clin Pathol 2016 Jul 22:jclinpath-2016-203872. [doi: 10.1136/jclinpath-2016-203872] [Medline: 27451435]
49. Eppenga WL, Derijks HJ, Conemans JMH, Hermens W, Wensing M, De Smet PAGM. Comparison of a basic and an advanced pharmacotherapy-related clinical decision support system in a hospital care setting in the Netherlands. J Am Med Inform Assoc 2012;19(1):66-71. [doi: 10.1136/amiajnl-2011-000360] [Medline: 21890873]

Abbreviations

ADE: adverse drug event
C3RG: Central Catalonia Chronicity Research Group
CDSS: clinical decision support system
CPOE: computerized physician order entry

DBI: Drug Burden Index
DSS: Decision support system
EHR: electronic health record
LESS-CHRON: List of Evidence-Based Deprescribing for Chronic Patients
MRCI: Medication Regimen Complexity Index
MRCI-E: Spanish Medication Regimen Complexity Index
OWL: Web Ontology Language
STOPP/START: Screening Tool of Older Persons' Prescriptions and Screening Tool to Alert to Right Treatment

Edited by C Perrin; submitted 20.01.23; peer-reviewed by J Cimino, T Salzmann, V Gadicherla; revised version received 30.03.23; accepted 03.04.23; published 10.07.23.

Please cite as:
Calvo-Cidoncha E, Verdinelli J, González-Bueno J, López-Soto A, Camacho Hernando C, Pastor-Duran X, Codina-Jané C, Lozano-Rubí R
An Ontology-Based Approach to Improving Medication Appropriateness in Older Patients: Algorithm Development and Validation Study
JMIR Med Inform 2023;11:e45850
URL: https://medinform.jmir.org/2023/1/e45850
doi:10.2196/45850

Original Paper

Agreement Between Experts and an Untrained Crowd for Identifying Dermoscopic Features Using a Gamified App: Reader Feasibility Study

Jonathan Kentley[1,2*], MSc, MBBS; Jochen Weber[2*]; Konstantinos Liopyris[3], MD, PhD; Ralph P Braun[4], MD; Ashfaq A Marghoob[2], MD; Elizabeth A Quigley[2], MD; Kelly Nelson[5], MD; Kira Prentice[6], BA; Erik Duhaime[6], MPhil, PhD; Allan C Halpern[2], MD; Veronica Rotemberg[2], MD, PhD

[1]Department of Dermatology, Chelsea and Westminster Hospital, London, United Kingdom

[2]Dermatology Section, Memorial Sloan Kettering Cancer Center, New York, NY, United States

[3]Department of Dermatology, Andreas Syggros Hospital of Cutaneous and Venereal Diseases, University of Athens, Athens, Greece

[4]Department of Dermatology, University Hospital Zurich, Zurich, Switzerland

[5]Department of Dermatology, The University of Texas MD Anderson Cancer Center, Houston, TX, United States

[6]Centaur Labs, Boston, MA, United States

[*]these authors contributed equally

Corresponding Author:
Veronica Rotemberg, MD, PhD
Dermatology Section
Memorial Sloan Kettering Cancer Center
530 E 74th Street
New York, NY, 10021
United States
Phone: 1 8336854126
Email: rotembev@mskcc.org

Abstract

Background: Dermoscopy is commonly used for the evaluation of pigmented lesions, but agreement between experts for identification of dermoscopic structures is known to be relatively poor. Expert labeling of medical data is a bottleneck in the development of machine learning (ML) tools, and crowdsourcing has been demonstrated as a cost- and time-efficient method for the annotation of medical images.

Objective: The aim of this study is to demonstrate that crowdsourcing can be used to label basic dermoscopic structures from images of pigmented lesions with similar reliability to a group of experts.

Methods: First, we obtained labels of 248 images of melanocytic lesions with 31 dermoscopic "subfeatures" labeled by 20 dermoscopy experts. These were then collapsed into 6 dermoscopic "superfeatures" based on structural similarity, due to low interrater reliability (IRR): dots, globules, lines, network structures, regression structures, and vessels. These images were then used as the gold standard for the crowd study. The commercial platform DiagnosUs was used to obtain annotations from a nonexpert crowd for the presence or absence of the 6 superfeatures in each of the 248 images. We replicated this methodology with a group of 7 dermatologists to allow direct comparison with the nonexpert crowd. The Cohen κ value was used to measure agreement across raters.

Results: In total, we obtained 139,731 ratings of the 6 dermoscopic superfeatures from the crowd. There was relatively lower agreement for the identification of dots and globules (the median κ values were 0.526 and 0.395, respectively), whereas network structures and vessels showed the highest agreement (the median κ values were 0.581 and 0.798, respectively). This pattern was also seen among the expert raters, who had median κ values of 0.483 and 0.517 for dots and globules, respectively, and 0.758 and 0.790 for network structures and vessels. The median κ values between nonexperts and thresholded average–expert readers were 0.709 for dots, 0.719 for globules, 0.714 for lines, 0.838 for network structures, 0.818 for regression structures, and 0.728 for vessels.

Conclusions: This study confirmed that IRR for different dermoscopic features varied among a group of experts; a similar pattern was observed in a nonexpert crowd. There was good or excellent agreement for each of the 6 superfeatures between the

crowd and the experts, highlighting the similar reliability of the crowd for labeling dermoscopic images. This confirms the feasibility and dependability of using crowdsourcing as a scalable solution to annotate large sets of dermoscopic images, with several potential clinical and educational applications, including the development of novel, explainable ML tools.

(JMIR Med Inform 2023;11:e38412) doi:10.2196/38412

KEYWORDS

dermatology; dermatologist; diagnosis; diagnostic; labeling; classification; deep learning; dermoscopy; dermatoscopy; skin; pigmentation; microscopy; dermascopic; artificial intelligence; machine learning; crowdsourcing; crowdsourced; melanoma; cancer; lesion; medical image; imaging; development; feasibility

Introduction

The use of dermoscopy, a low-cost, noninvasive diagnostic technique based on a hand-held device with a light source and magnifying lens, is routine practice for the evaluation of pigmented skin lesions and has been shown to increase sensitivity for early melanoma detection [1,2]. Dermoscopy allows examination of morphological features below the stratum corneum that would not be visible by visual inspection alone [3]. Diagnosis of melanoma using dermoscopy relies on assessment of lesion morphology and identification of dermoscopic features. A number of diagnostic criteria and algorithms have been developed for this purpose, including pattern analysis [4], the ABCD (asymmetry, border, color, diameter) rule [5], the Menzies method [6], the 7-point checklist [7], and the CASH (color, architecture, symmetry, homogeneity) score [8].

As use of dermoscopy has expanded, so too has dermoscopic vocabulary, resulting in a vast number of published feature definitions and 2 competing terminologies: metaphoric and descriptive. In recent years, efforts have been made to harmonize nomenclature, and the 2016 International Dermoscopy Society terminology consensus proposed 31 specific “subfeatures” of melanocytic lesions, falling into 9 “superfeatures” based on structural similarities (Textbox 1) [9].

However, interrater reliability (IRR) for identifying melanoma-specific dermoscopic structures has been shown to be poor [10]. Our research group recently performed the EASY (Expert Agreement on the Presence and Spatial Location of Melanocytic Features in Dermoscopy) study, which found that agreement was highly variable when 20 dermoscopy experts were asked to identify the 31 dermoscopic subfeatures in an image set specifically curated for this purpose. IRR across 248 images was poor to moderate for all but 7 features. We demonstrated that when individual subfeatures were collapsed into 9 superfeatures, increased agreement was observed, ranging from a pairwise Fleiss κ of 0.14 for the detection of dots to 1.0 for the detection of a pigment network structure.

Machine learning (ML) methods have recently been investigated in the field of dermatology, and the majority of developed algorithms are diagnostic binary classifiers [11,12]. A number of studies have evaluated the performance of algorithms developed to detect specific dermoscopic features, including pigment network structures, vessels, and blue-white veil; however, many algorithms were trained and tested on relatively small data sets and have achieved only moderate accuracy [13-21].

Due to the vast dimensionality of medical images, classifier algorithms are typically of an uninterpretable “black box” nature, a term that describes the phenomenon whereby functions that connect input pixel data to output labels cannot be understood by the human brain. There has been a push by medical regulators and the artificial intelligence community to develop explainable algorithms; however, it has been acknowledged that this may come at the cost of decreased accuracy [22]. Incorporating detection of dermoscopic features into melanoma classifier algorithms may allow for better explainability and therefore greater acceptance into clinical practice by clinicians and regulatory bodies [23,24].

The International Skin Imaging Collaboration (ISIC) archive provides an open-access data set comprising almost 70,000 publicly available dermoscopic images at the time of writing, including 5598 melanomas and 27,878 nevi. As well as hosting the regular ISIC Grand Challenge to promote the development of ML for melanoma detection, the archive has been extensively utilized to train independent ML algorithms and acts as a comprehensive educational resource for dermatologists via the Dermoscopedia platform [25,26]. Most public images in the archive have labels serving as a diagnostic ground truth for supervised learning. However, accurate feature annotations are thus far lacking. As part of the 2018 ISIC Challenge, 2595 images were annotated for 5 dermoscopic patterns (pigment network structures, negative network structures, streaks, milia-like cysts, and dots/globules) [27]. However, the ground truth labels were provided by only 1 clinician and the performance of the 23 submitted algorithms was acknowledged to be exceptionally low, likely as a result of this [27].

As medical data sets continue to rapidly expand and computing power increases, it is widely recognized that one of the major limiting factors for the development of robust and generalizable ML in dermatology is the need for large, comprehensively labeled data sets [28,29]. Obtaining annotations of medical images by medical experts is both time-consuming and expensive, creating a bottleneck in the development pipeline and making it challenging to obtain annotations at scale [30].

Crowdsourcing provides a potential solution to these problems. Crowdsourcing involves the recruitment of groups of individuals of varying levels of knowledge, heterogeneity, and number who voluntarily complete an online task, often with financial incentives [31,32]. Monetary compensation is typically less than US $0.10 per annotation, and tasks can be distributed to a

large number of workers in parallel, aggregating the crowd's knowledge to complete the task in a cost- and time-effective manner [33,34]. One study reported that it took 6 months to obtain expert labels comprising 340 sentences from radiology reports written by 2 radiologists, whereas the authors obtained crowdsourced annotations of 717 sentences in under 2 days at a cost of less than $600. A classification algorithm trained using these crowdsourced annotations outperformed an algorithm trained using the expert-labeled data as a result of the increased volume of available training examples [32].

Given the heterogeneity of biomedical data, the utility of crowdsourcing may decrease with the complexity of the task. For example, the 14 million images contained in the ImageNet archive were easily annotated by the untrained public, whereas the ability to classify and segment radiological images may require many years of specialist training [28,30,35]. Nevertheless, crowdsourcing has proven effective in a wide range of applications for biomedical imaging, most commonly histopathology or retinal imaging [34].

Feng et al [36] reported that a crowd of South Korean students were able to reach similar diagnostic accuracy as experts for diagnosing malaria-infected red blood cells after only 3 hours of training, allowing the authors to build a gold standard library of malaria-infection labels for erythrocytes. The authors used a game-based tool that made the task easy to complete by including points and a leaderboard on the platform. This method of so-called gamification is frequently used by crowdsourcing platforms and has been shown to increase the engagement of the crowd and improve the quality of the crowdsourced work [37]. Bittel et al [38] used a hybrid crowd-ML approach to create the largest publicly available data set of annotated endoscopic images. Heim et al [28] found that a crowd was able to segment abdominal organs in computed tomography (CT) images with comparable quality to a radiologist, but at a rate up to 350 times faster.

There are few studies published to date evaluating crowdsourcing in the field of dermatology, and to the best of the authors' knowledge, there are no published studies on the utility of crowdsourcing for the annotation of features present in dermoscopic images [39,40].

The aim of this study is to demonstrate that crowdsourcing can be employed to label dermoscopic subfeatures of melanocytic lesions with equivalent reliability to a small group of dermatologists. This will allow for efficient annotation of a large repository of dermoscopic images to aid the development of novel ML algorithms [32]. Incorporating detection of dermoscopic features into diagnostic algorithms will result in explainable outputs and may therefore improve the acceptability of these outputs to the medical community.

Textbox 1. List of superfeatures (in bold) and corresponding subfeatures seen in melanocytic lesions [9].

Dots

Irregular, regular

Globules

Cobblestone pattern, irregular, regular, rim of brown globules

Lines

Branched streaks, pseudopods, radial streaming, starburst

Network structures

Atypical pigment network, broadened pigment network, delicate pigment network, negative pigment network, typical pigment network

Regression structures

Peppering/granularity, scarlike depigmentation

Shiny white structures

Patterns

Angulated lines, polygons, zigzags

Structureless areas

Irregular blotches, regular blotches, blue-whitish veil, milky red areas, structureless brown areas, and homogenous (not otherwise specified)

Vessels

Comma, corkscrew, dotted vessel, linear irregular vessel, polymorphous vessel, milky red globules

Methods

Ethics Approval

This study was conducted as part of the umbrella ISIC research protocol and was approved by the Memorial Sloan Kettering Cancer Center Institutional Review Board (16-974). All images were deidentified and do not contain any protected health information as per the terms of use agreement for the ISIC archive.

Materials

This study was performed in 3 separate experiments, each using the same set of 248 lesion images used in the EASY study. Briefly summarized, clinical experts contributed 964 lesion images showing 1 of 31 preselected subfeatures, as described

by Kittler et al [9]. Clinicians were asked to submit images of "excellent quality showing the exemplar feature in focus." Three experts chose 248 of these images, roughly balancing benign and malignant lesions and ensuring image quality. Each of the 31 features was the exemplar in 8 of the lesion images submitted. However, each image could, and typically did, show multiple features.

Subfeatures and Superfeatures

As described earlier, low to moderate IRR was observed for the majority of subfeatures. Hence, we used only the superfeature terms for our scalability investigation. While each of the subfeatures had 8 exemplar images, collapsing the labels into superfeatures created some imbalance. The full list of subfeatures is shown in Textbox 1. The 9 superfeatures (dots, globules, lines, network structures, patterns, regression structures, shiny white structures, structureless areas, and vessels; shown in Multimedia Appendix 1, Table S1) were presented to participants during the tutorial on the DiagnosUs smartphone app, adapted from Marghoob and Braun [41].

Agreement Measure

To measure agreement across raters, we employed the Cohen κ [42], which has a value of 0 for completely random choices, increasing toward a maximum value of 1.0 with improved IRR. Measures of agreement are interpreted as poor (0-0.4), fair to good (≥0.4-0.75), and excellent (≥0.75-1.0) [43]. This measure was primarily chosen to accommodate the nature of the 3 separate studies (see below), allowing for partial data between pairs of raters using the binary choice of "feature present" or "feature absent." Throughout this paper, we use the term "median κ" to refer to the median of κ values across the set of pairwise comparisons as a measure of central tendency, given the nonnormal distribution of κ values.

Initial Expert Annotations (Study 1)

For the first study, we used a custom programmed annotation platform built for the ISIC archive. We asked a total of 20 dermoscopy experts to each annotate 62 images (2 per exemplar feature) in 4 substudies of nonoverlapping image sets. Experts for study 1 were clinicians with ≥10 years of dermoscopy experience who had made significant contributions to dermoscopy research or teaching dermoscopy of pigmented lesions. For each image, 5 experts were asked to provide benign/malignant status and then to self-select which of the 31 available subfeatures they perceived as present in the image. Full data and results of the EASY study will be published separately.

Gold Standard for the Crowd Study

After collapsing the subfeatures into the 9 abovementioned superfeatures, we found that 3 had very poor agreement and too few exemplars to allow reliable evaluation by the crowd: patterns, shiny white structures, and structureless areas. For the remaining 6 superfeatures (dots, globules, lines, network structures, regression structures, and vessels), images in which at least 3 of 5 experts in study 1 had selected ≥1 of the subfeatures within the same superfeature as present were used as the gold standard for "superfeature present." Images in which none of the 5 experts had identified any of the subfeatures within the same superfeature as present were used as the gold standard for "superfeature absent."

Nonexpert Crowd Annotations (Study 2)

To collect nonexpert image annotations, we used the commercially available platform DiagnosUs (Centaur Labs) [44] through a collaboration agreement. Users can sign up to the app and participate in competitions, which increases engagement and improves accuracy [37]. Users are recruited via a referral system or advertisements on social media. To ensure that only users somewhat skilled at a task computed average detection values, gold standard images were used for both training and validation. This left the remaining images, for which either 1 or 2 expert raters annotated a subfeature within the same superfeature as being present, as true test images. If a user did not reach at least 83% correctness for the validation items, that user's choices were not used in the subsequent analysis. Each of the 6 superfeatures was presented as a separate task. In addition to the binary choice of presence or absence of a superfeature, we also collected reaction times to assess decision difficulty [45].

Expert Crowd Annotations (Study 3)

As study 1 allowed experts to select from the 31 subfeatures, we replicated the methodology of study 2 to allow direct comparison with the nonexpert crowd. Experts in study 2 were dermatologists with ≥5 years of experience. We recruited 7 experts to use the DiagnosUs platform and annotate the same 248 images from studies 1 and 2 for the presence of the 6 superfeatures. For each of the features, we selected the first 5 dermatologists who completed annotation of the image set.

Reaction Times

For each of the tasks in studies 2 and 3, we computed the per-item averaged logged reaction times as the log of (1 + reaction time) to approximate a normal distribution of measurement errors. These averaged logged reaction times were then regressed against the average responses and a quadratic term, allowing for an inverted-U–shaped response function, which peaked roughly at the (across-readers) point of indecision.

Results

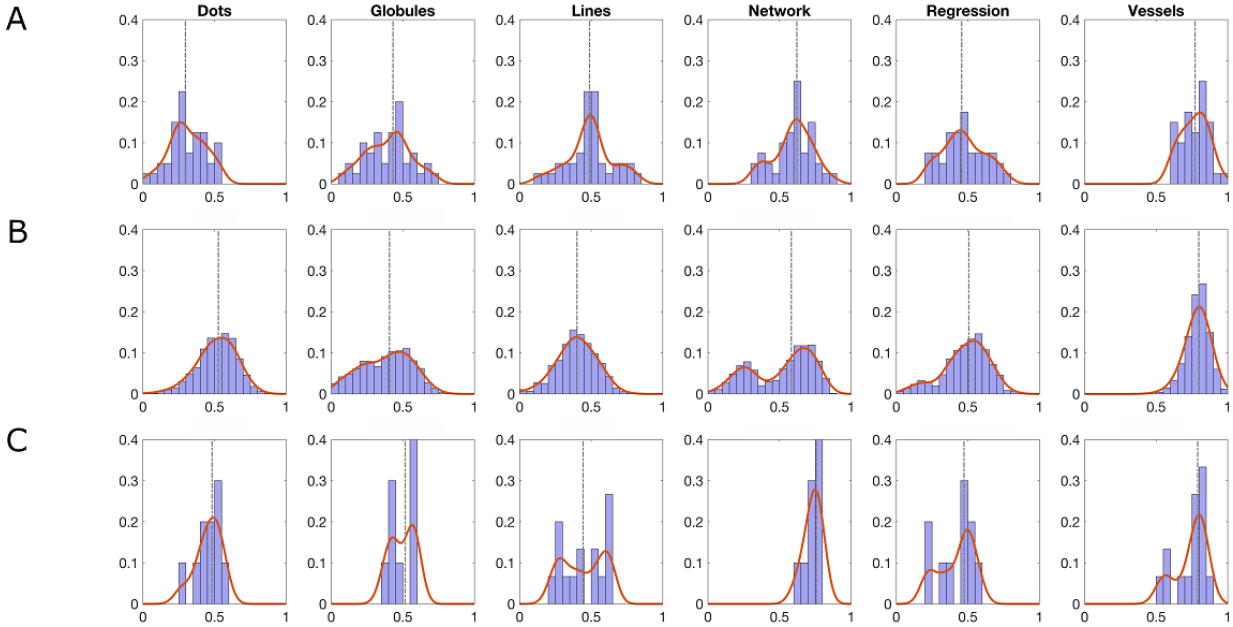
Initial Expert Annotations (Study 1)

In study 1, we found that dots showed poor agreement (median κ=0.298), whereas vessels showed excellent agreement (median κ=0.768). All other superfeatures showed fair to good agreement (Table 1). The resulting distributions of pairwise Cohen κ values are shown in Figure 1A. The number of resulting gold-standard images for each of the 6 superfeatures was as follows (0 readers/at least 3 readers, respectively): dots (93/61), globules (57/92), lines (129/60), network structures (63/140), regression structures (113/59), and vessels (152/66).

Table 1. Median Cohen κ values for pairwise readers. For study 2, pairs of readers were considered only if both readers saw at least 62 of the same images.

Feature	Study 1 (experts), median κ	Study 2 (nonexpert crowd), median κ	Study 3 (expert crowd), median κ
Dots	0.2977	0.5264	0.4829
Globules	0.4075	0.3945	0.5166
Lines	0.5205	0.3983	0.4433
Network structures	0.6175	0.5810	0.7575
Regression structures	0.4643	0.5066	0.4730
Vessels	0.7683	0.7977	0.7903

Figure 1. Pair-wise Cohen κ values for study 1 (A), study 2 (B), and study 3 (C).



Nonexpert Crowd Annotations (Study 2)

Providing demographic data pertaining to the users' jobs and their reasons for using the DiagnosUs platform was optional; these data were collected from 190 users. Of these, 23 (12.1%) were physicians (2 dermatologists, 21 other specialties), 72 (37.9%) were medical students, 11 (5.8%) were nurse practitioners, 8 (4.2%) were physician assistants, and 76 (40%) were "other" or "other healthcare student." The most common reason for using DiagnosUs was "improve my skills" (134/190, 70.5%), followed by "earn money" (37/190, 19.5%) and "compete with others" (19/190, 10%).

The number of users that engaged with each of the features varied for dots (92 users), globules (111 users), lines (82 users), network structures (97 users), regression structures (79 users), and vessels (95 users). Equally, the median number of ratings made per user per task varied for dots (160 images rated per user), globules (131 images), lines (177 images), network structures (91 images), regression structures (124 images), and vessels (104 images). The total number of crowd base ratings obtained in this study was 139,731, including 25,466 total ratings for dots, 40,853 for globules, 21,074 for lines, 17,114 for network structures, 17,020 for regression structures, and 18,204 for vessels.

The pattern we found in study 1 was largely replicated by the nonexperts. To ensure that there was sufficient and comparable overlap for images between pairs of readers, only pairs in which both readers saw at least 62 of the same images were evaluated. Dots and globules showed relatively lower agreement (with median κ values of 0.526 and 0.395, respectively), whereas network structures and vessels showed the highest agreement (with median κ values of 0.581 and 0.798, respectively). To allow a direct comparison between studies 1 and 2, we have compiled the 6 superfeatures into a panel figure (Figure 1A and 1B).

Expert Crowd Annotations (Study 3)

Again, the patterns found in studies 1 and 2 were replicated, such that dots and globules showed relatively lower agreement (median κ values were 0.483 and 0.517, respectively), whereas network structures and vessels showed the highest agreement (median κ values were 0.758 and 0.790, respectively; Figure 1C).

We computed κ values for each nonexpert reader from study 2 compared to a single simulated expert by thresholding the responses in study 3 from 3 of 5 experts into a binary variable. The median κ values between nonexperts and the thresholded average expert reader were as follows: for dots, 0.709; for

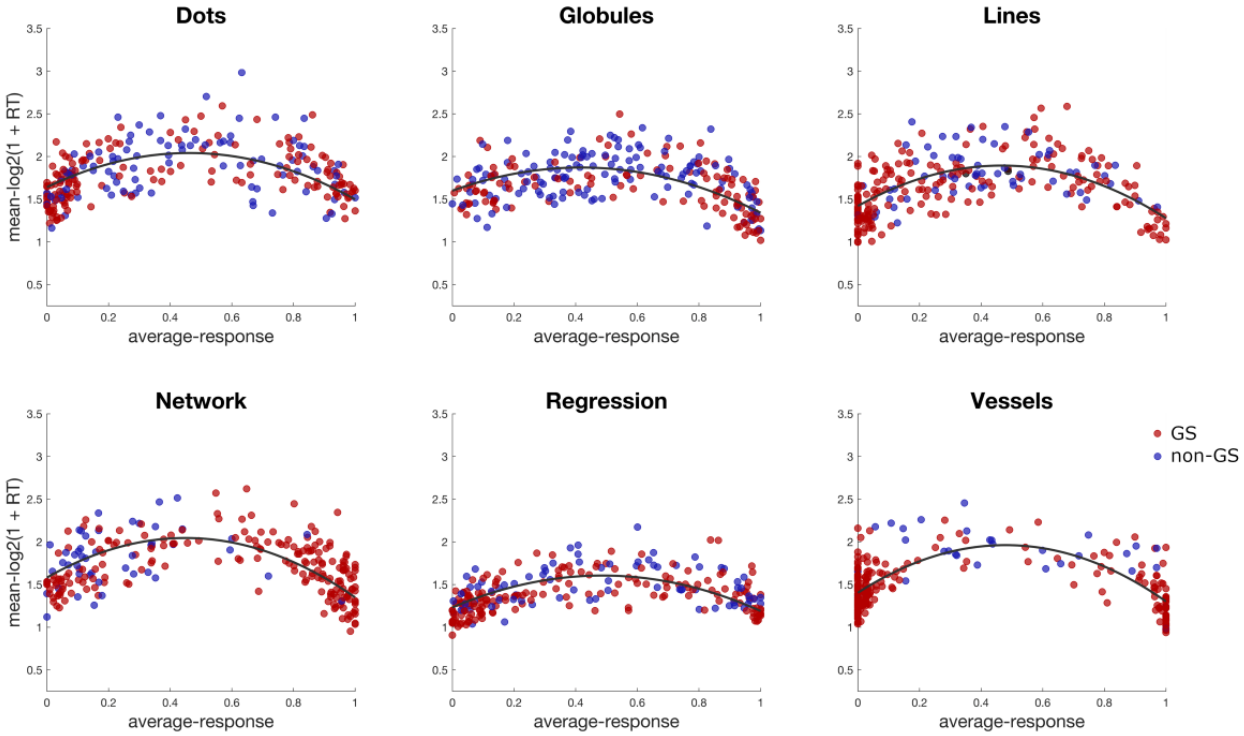
globules, 0.719; for lines, 0.714; for network structures, 0.838; for regression structures, 0.818; and for vessels, 0.728.

Reaction Times

Irrespective of task, the reaction time varied by user (the median IQR for reaction time across users was 2.5 seconds to 4.3 seconds) and across images (the median IQR for the difference in reaction time per user was –0.93 seconds to +1.5 seconds), suggesting that the variability within users was somewhat greater than the variability across users.

For both the nonexperts and experts, the quadratic term accounting for the inverted-U–shaped response in averaged logged reaction times reached statistical significance across all tasks. Among the nonexperts, the *t* values (calculated with a 2-tailed *t* test) ranged from t_{244}=–14.3 (for dots) to t_{244}=–20.09 (for vessels). Among the experts, probably due to higher noise, the *t* values ranged from t_{244}=–7.63 (for regression structures) to t_{244}=–10.62 (for vessels). All *t* values were highly significant (*P*<.001). In all tasks and for both sets of readers, the linear term had a negative sign and was also significant (at lower levels), meaning that in all cases readers were faster to respond when a feature was present compared to when it was absent (Figure 2).

Figure 2. Log reaction times for gold standard images (shown by the red dots) and non–gold standard images (shown by the blue dots) of nonexperts regressed against their average responses and showing the estimated quadratic term for each superfeature. RT: reaction time; GS: gold standard.



Discussion

Principal Findings

The main findings of this study confirmed the variable, and sometimes low, IRR between experts for identifying dermoscopic superfeatures on images of melanocytic lesions. The patterns of repeatability were mirrored in all 3 studies, highlighting that some features are more challenging to identify regardless of experience level. We found that the IRR between the untrained crowd and expert crowd was good to excellent for all superfeatures, suggesting that crowdsourced labels can be reliably used for future research. Reaction times were slower for lesions that would be considered more challenging in both cohorts, and therefore may be used as a proxy for decision difficulty.

Initial Expert Annotations (Study 1)

In Study 1, the lowest level of agreement was observed for dots and globules, and the highest agreement was observed for network structures and vessels. This is in keeping with the findings of previous studies evaluating IRR for the identification of dermoscopic patterns among a group of experienced dermoscopists [10,46]. It has been suggested that poor agreement on criteria such as structureless areas, streaks, and dots or globules may be the result of lack of standardization in dermoscopy education [46,47].

Furthermore, the definition of dermoscopic structures may evolve over time. Whereas vascular structures and pigment network structures are easily recognizable, and their definitions have been consistent in the literature to date, dots and globules may be less easy to categorize. Tiny, numerous gray dots may be categorized as regression structures, and red dots may be defined as vascular structures [48-50]. Globules are defined as measuring >0.1 mm, which may be challenging to identify in dermoscopic images without a unit of measurement as a reference point. Going forward, it may be more feasible to consider dots and globules as a single criterion to eliminate the

challenges encountered when attempting to differentiate them based on size.

Nonexpert Crowd Annotations (Study 2)

A similar pattern of results was seen in study 2, suggesting that the gridlike pattern of a pigment network structure and the distinctive red color of vascular structures may be more repeatably identified by an untrained crowd. In keeping with the results of study 1, dots and globules were identified with poor repeatability. Again, this may be as a result of the ambiguity in distinguishing between the two on the basis of their diameter.

Prior studies have shown that dermoscopy by novice clinicians is no more accurate than visual inspection alone, and so an untrained crowd would not be expected to identify complex dermoscopic patterns, particularly when agreement between a group of world experts is known to be low, such as in our EASY study. To obtain reliable crowdsourced labels for complex medical images, an easier set of images may be used or participants may receive extended training; the study must also be designed to accommodate a large number of redundant labels [28]. In a study evaluating crowdsourcing as a method of identifying colonic polyps in CT colonoscopy images, McKenna et al [51] found that the crowd performance deteriorated with increasing difficulty, as well as with increasing reaction time. By collapsing the 31 subfeatures into 6 superfeatures, we created a more achievable task for a crowd with no prior experience of dermoscopy.

Expert Crowd Annotations (Study 3)

The results from study 3 showed that agreement between experts was higher for dots, globules, and network structures when compared to study 1, in which annotations for subfeatures were aggregated into superfeature categories. It is known that there is a greater potential for disagreement with an increased number of categories and that the Cohen κ is typically observed to be lower in this circumstance [52]. Thus, if experts has been asked to choose from 6 superfeatures rather than 31 subfeatures, there would have been less potential for disagreement.

When comparing the median κ across all 3 studies, we found that repeatability for identifying all 6 superfeatures was similar across the experts and nonexperts. When comparing the median nonexpert annotations in study 2 to the thresholded expert annotations in study 3 for the same task, we saw that agreement was excellent for network structures and regression structures and good for the 4 remaining superfeatures. This suggests that the crowd was able to both repeatably and reliably identify dermoscopic superfeatures. Interestingly, agreement for vessels was higher within groups than between groups; thus, crowd annotations, although repeatable, were less accurate than expert annotations, suggesting that the crowd may be less reliable when annotating vessels. Vessels had the highest number of subfeatures (6) with distinct morphologies, several of which were not presented to the crowd during training on the DiagnosUs platform. Redesigning the tutorial may result in better accuracy for crowd annotations of vessels.

Reaction Times

For both experts and nonexperts, there were 2 common patterns of response time (ie, the time it took a participant to feel confident enough to log a response varied as a function of estimated difficulty). For images for which the crowd showed low agreement (the average response was approximately 0.5 seconds), the response times were significantly slower than for images for which the crowd showed high agreement. For gold standard images (those for which ≥3 of 5 experts in study 1 agreed on the presence or absence of a feature) reaction times were faster than those for images of lesions upon which only 1 or 2 experts agreed, highlighting the challenging nature of these images. Furthermore, images where the feature was present had faster reaction times than those where the feature was absent, regardless of level of agreement. Overall, experts took longer to respond to images than nonexperts, suggesting that they exerted more effort to ensure a correct response. In addition, there was no financial reward for experts in this study; thus, they were less motivated to annotate as many lesions as possible within a designated timeframe.

Limitations

One of the fundamental limitations of this study and future implications that can be drawn from it is the potentially low dependability of crowdsourced annotations. Although we found high repeatability and reliability of labels in study 3, this was for a relatively small set of images that had been carefully curated to have high-quality examples of a limited number of superfeatures.

There are a number of proposed methods to improve the quality of crowdsourced data. Crowd performance has been shown to improve with increased time spent training for the task, and participants that complete more readings have been observed to perform better [36,53]. Therefore, we may be able to improve performance of the crowd by providing additional training, as well as by increasing participant engagement, such as with greater financial rewards. This may, however, come at the cost of increased time and a smaller number of participants. Although crowdsourced annotations may be marginally less accurate than those provided by experts, the increased number of available labels for training ML algorithms has been shown to make them more robust to noisy data [54].

In this study, we validated the participants' performance against gold standard images to ensure the quality of labels, and poorly performing participants were not included. In the absence of an expert-labeled image, DiagnosUs allows a ground truth to emerge with an unlabeled competition design in which images that show internal consistency across raters become the initial gold standard. Filtering of individuals may also be achieved by evaluating participants based on previously performed tasks or providing a pretask test [34]. Aggregating results via majority voting is another commonly used method of preprocessing to improve annotation quality. Annotations may also be evaluated by using them to train a ML model and using the model's performance as a proxy for crowd performance [34].

It is essential that some level of quality assurance take place for crowdsourced annotations in the absence of expert labeling for

comparison, as would be the case in future studies. Although agreement is traditionally considered an indicator of data reliability, it has been suggested that participants' competence and confidence should be taken into account [55]. This can be achieved by filtering participants with poor accuracy on gold standard images, aggregating annotations, and using reaction time as a proxy for decision confidence. Images that give rise to long reaction times and a low level of agreement may then be transferred to an expert for annotation.

Many of the lesions in the archive are complex and have multiple dermoscopic patterns, which we observed created challenges for the experts to reliably identify, let alone the untrained crowd. Obtaining annotations for only 6 superfeatures may limit the diagnostic value of an ML tool. Crowdsourced labeling of the ISIC archive may be limited by its size; at the time this study was conducted, approximately 10,000 superfeature annotations were collected per day. However, engagement with the DiagnosUs platform continues to grow exponentially, and it currently receives in excess of 1 million crowd opinions daily across multiple tasks. Therefore, it may be entirely achievable to annotate the ISIC archive with crowdsourced labels within a timeframe of weeks to months.

Although the images in this study were subject to a manual quality assurance process, they were not standardized. For example, some images contained a unit of measure, which may have introduced bias when differentiating between dots and globules, as mentioned earlier in the discussion.

Insufficient demographic data were collected by the DiagnosUs platform to allow meaningful subanalyses; however, disparities in experience level between users were highlighted. Importantly, 2 physicians specializing in dermatology participated in the crowd, and it therefore cannot be truly considered untrained. Due to the nature of the platform, it appeals to medical professionals as a learning tool with the aim of driving innovation in medical artificial intelligence, and the platform provides meaningful labels at scale regardless of the background of its users.

Future Work

Given the sheer size of the ISIC archive, it would be infeasible to obtain annotations by expert dermoscopists for all images. We have shown the feasibility of obtaining crowdsourced annotations; this method can be used in several ways. First, it will allow hierarchical organization of the archive, allowing users to filter lesions based on dermoscopic patterns. Second, it may act as a teaching tool, allowing novice dermoscopists to learn patterns and corresponding diagnoses. And third, these annotated data may be used to develop novel ML tools. Even if only a small proportion of images are labeled by the crowd, a pattern classification or segmentation algorithm could be used to annotate additional images in the archive though a weakly supervised technique [56]. A hybrid crowd-algorithm approach has been successfully developed by several groups for the purpose of segmenting large databases of medical images [28,38,54,57].

The issues regarding "black box" algorithms have been raised as a barrier to implementation of these tools in clinical practice. Given the complexity of medical imaging data, a fully explainable algorithm would be unlikely to have adequate performance; however, use of interpretable outputs may go some way to assuage hesitancy in uptake. A classification tool that is also able to detect dermoscopic patterns that have influenced its decision would allow dermatologists to make more informed decisions when evaluating the output of the algorithm [22]. Furthermore, a multidimensional algorithm that is trained on both diagnoses and dermoscopic features may have increased accuracy when compared to those trained on diagnoses alone.

The next steps in exploring the applications of crowdsourced data are to expand labeling to a larger sample of images with a robust quality assurance process and incorporate the labels into a pattern-detection algorithm to be evaluated in a study of readers. Should this algorithm display acceptable performance measures, it may be deployed to label further images and be incorporated into a classification algorithm to improve its explainability.

Conflicts of Interest

JK has provided services for Skin Analytics, Ltd and IQVIA, Inc. ACH has provided services for Canfield Scientific, Lloyd Charitable Trust, and SciBase; has ownership and equity interests in HCW Health, LLC; and has a fiduciary role, intellectual property rights, and ownership and equity interests in SKIP Derm, LLC. VR has provided services for Inhabit Brands, Ltd. AAM has received royalties from UpToDate. KN is an employee of Centaur Labs. KP is an employee of Centaur Labs. ED is the CEO of Centaur Labs.

Multimedia Appendix 1
Select dermoscopic features as presented to participants during the tutorial on the DiagnosUs smartphone app.
[DOCX File , 2640 KB - medinform_v11i1e38412_app1.docx]

References

1. Murzaku EC, Hayan S, Rao BK. Methods and rates of dermoscopy usage: a cross-sectional survey of US dermatologists stratified by years in practice. J Am Acad Dermatol 2014 Aug;71(2):393-395. [doi: 10.1016/j.jaad.2014.03.048] [Medline: 25037790]

2. Dinnes J, Deeks JJ, Chuchu N, Ferrante di Ruffano L, Matin RN, Thomson DR, Cochrane Skin Cancer Diagnostic Test Accuracy Group. Dermoscopy, with and without visual inspection, for diagnosing melanoma in adults. Cochrane Database Syst Rev 2018 Dec 04;12(12):CD011902 [FREE Full text] [doi: 10.1002/14651858.CD011902.pub2] [Medline: 30521682]
3. Celebi ME, Codella N, Halpern A. Dermoscopy image analysis: overview and future directions. IEEE J Biomed Health Inform 2019 Mar;23(2):474-478. [doi: 10.1109/JBHI.2019.2895803] [Medline: 30703051]
4. Carli P, Quercioli E, Sestini S, Stante M, Ricci L, Brunasso G, et al. Pattern analysis, not simplified algorithms, is the most reliable method for teaching dermoscopy for melanoma diagnosis to residents in dermatology. Br J Dermatol 2003 May;148(5):981-984. [doi: 10.1046/j.1365-2133.2003.05023.x] [Medline: 12786829]
5. Nachbar F, Stolz W, Merkle T, Cognetta AB, Vogt T, Landthaler M, et al. The ABCD rule of dermatoscopy. High prospective value in the diagnosis of doubtful melanocytic skin lesions. J Am Acad Dermatol 1994 Apr;30(4):551-559. [doi: 10.1016/s0190-9622(94)70061-3] [Medline: 8157780]
6. Menzies SW, Ingvar C, Crotty KA, McCarthy WH. Frequency and morphologic characteristics of invasive melanomas lacking specific surface microscopic features. Arch Dermatol 1996 Oct;132(10):1178-1182. [Medline: 8859028]
7. Argenziano G, Fabbrocini G, Carli P, De Giorgi V, Sammarco E, Delfino M. Epiluminescence microscopy for the diagnosis of doubtful melanocytic skin lesions. Comparison of the ABCD rule of dermatoscopy and a new 7-point checklist based on pattern analysis. Arch Dermatol 1998 Dec;134(12):1563-1570. [doi: 10.1001/archderm.134.12.1563] [Medline: 9875194]
8. Henning JS, Stein JA, Yeung J, Dusza SW, Marghoob AA, Rabinovitz HS, et al. CASH algorithm for dermoscopy revisited. Arch Dermatol 2008 Apr;144(4):554-555. [doi: 10.1001/archderm.144.4.554] [Medline: 18427058]
9. Kittler H, Marghoob AA, Argenziano G, Carrera C, Curiel-Lewandrowski C, Hofmann-Wellenhof R, et al. Standardization of terminology in dermoscopy/dermatoscopy: Results of the third consensus conference of the International Society of Dermoscopy. J Am Acad Dermatol 2016 Jun;74(6):1093-1106 [FREE Full text] [doi: 10.1016/j.jaad.2015.12.038] [Medline: 26896294]
10. Carrera C, Marchetti MA, Dusza SW, Argenziano G, Braun RP, Halpern AC, et al. Validity and reliability of dermoscopic criteria used to differentiate nevi from melanoma: a web-based International Dermoscopy Society study. JAMA Dermatol 2016 Jul 01;152(7):798-806 [FREE Full text] [doi: 10.1001/jamadermatol.2016.0624] [Medline: 27074267]
11. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017 Feb 02;542(7639):115-118 [FREE Full text] [doi: 10.1038/nature21056] [Medline: 28117445]
12. Thomsen K, Iversen L, Titlestad TL, Winther O. Systematic review of machine learning for diagnosis and prognosis in dermatology. J Dermatolog Treat 2020 Aug;31(5):496-510. [doi: 10.1080/09546634.2019.1682500] [Medline: 31625775]
13. Jaworek-Korjakowska J. A deep learning approach to vascular structure segmentation in dermoscopy colour images. Biomed Res Int 2018;2018:5049390 [FREE Full text] [doi: 10.1155/2018/5049390] [Medline: 30515404]
14. Kharazmi P, Zheng J, Lui H, Jane Wang Z, Lee TK. A computer-aided decision support system for detection and localization of cutaneous vasculature in dermoscopy images via deep feature learning. J Med Syst 2018 Jan 09;42(2):33. [doi: 10.1007/s10916-017-0885-2] [Medline: 29318397]
15. Demyanov S, Chakravorty R, Abedini M. Classification of dermoscopy patterns using deep convolutional neural networks. 2016 Presented at: 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI); Apr 13-16, 2016; Prague, Czech Republic p. 364. [doi: 10.1109/isbi.2016.7493284]
16. García Arroyo JL, García Zapirain B. Detection of pigment network in dermoscopy images using supervised machine learning and structural analysis. Comput Biol Med 2014 Jan;44:144-157 [FREE Full text] [doi: 10.1016/j.compbiomed.2013.11.002] [Medline: 24314859]
17. Anantha M, Moss RH, Stoecker WV. Detection of pigment network in dermatoscopy images using texture analysis. Comput Med Imaging Graph 2004 Jul;28(5):225-234 [FREE Full text] [doi: 10.1016/j.compmedimag.2004.04.002] [Medline: 15249068]
18. Sadeghi M, Razmara M, Wighton P, Lee TK, Atkins MS. Modeling the dermoscopic structure pigment network using a clinically inspired feature set. In: Liao H, Edwards PJ, Pan X, Fan Y, Yang GZ, editors. Medical Imaging and Augmented Reality. MIAR 2010. Lecture Notes in Computer Science, vol 6326. Berlin, Germany: Springer; 2010.
19. Maurya A, Stanley RJ, Lama N, Jagannathan S, Saeed D, Swinfard S, et al. A deep learning approach to detect blood vessels in basal cell carcinoma. Skin Res Technol 2022 Jul;28(4):571-576. [doi: 10.1111/srt.13150] [Medline: 35611797]
20. Madooei A, Drew MS, Sadeghi M, Atkins MS. Automatic detection of blue-white veil by discrete colour matching in dermoscopy images. In: Mori K, Sakuma I, Sato Y, Barillot C, Navab N, editors. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013. MICCAI 2013. Lecture Notes in Computer Science, vol 8151. Berlin, Germany: Springer; 2013.
21. Celebi ME, Iyatomi H, Stoecker WV, Moss RH, Rabinovitz HS, Argenziano G, et al. Automatic detection of blue-white veil and related structures in dermoscopy images. Comput Med Imaging Graph 2008 Dec;32(8):670-677 [FREE Full text] [doi: 10.1016/j.compmedimag.2008.08.003] [Medline: 18804955]
22. Babic B, Gerke S, Evgeniou T, Cohen IG. Beware explanations from AI in health care. Science 2021 Jul 16;373(6552):284-286. [doi: 10.1126/science.abg1834] [Medline: 34437144]

23. Mattessich S, Tassavor M, Swetter SM, Grant-Kels JM. How I learned to stop worrying and love machine learning. Clin Dermatol 2018;36(6):777-778. [doi: 10.1016/j.clindermatol.2018.06.003] [Medline: 30446202]
24. Price WN, Gerke S, Cohen IG. Potential liability for physicians using artificial intelligence. JAMA 2019 Nov 12;322(18):1765-1766. [doi: 10.1001/jama.2019.15064] [Medline: 31584609]
25. Rotemberg V, Halpern A, Dusza S, Codella NCF. The role of public challenges and data sets towards algorithm development, trust, and use in clinical practice. Semin Cutan Med Surg 2019 Mar 01;38(1):E38-E42. [doi: 10.12788/j.sder.2019.013] [Medline: 31051022]
26. Rotemberg V, Kurtansky N, Betz-Stablein B, Caffery L, Chousakos E, Codella N, et al. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. Sci Data 2021 Jan 28;8(1):34 [FREE Full text] [doi: 10.1038/s41597-021-00815-z] [Medline: 33510154]
27. Codella N, Gutman D, Emre Celebi M. Skin lesion analysis toward melanoma detection: A challenge at the 2017 International symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC). 2019 Presented at: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018); Apr 4-7, 2018; Washington, DC p. 3368. [doi: 10.1109/isbi.2018.8363547]
28. Heim E, Roß T, Seitel A, März K, Stieltjes B, Eisenmann M, et al. Large-scale medical image annotation with crowd-powered algorithms. J Med Imaging (Bellingham) 2018 Jul;5(3):034002 [FREE Full text] [doi: 10.1117/1.JMI.5.3.034002] [Medline: 30840724]
29. Park AJ, Ko JM, Swerlick RA. Crowdsourcing dermatology: DataDerm, big data analytics, and machine learning technology. J Am Acad Dermatol 2018 Mar;78(3):643-644. [doi: 10.1016/j.jaad.2017.08.053] [Medline: 29042152]
30. van der Wal D, Jhun I, Laklouk I, Nirschl J, Richer L, Rojansky R, et al. Biological data annotation via a human-augmenting AI-based labeling system. NPJ Digit Med 2021 Oct 07;4(1):145 [FREE Full text] [doi: 10.1038/s41746-021-00520-6] [Medline: 34620993]
31. Estellés-Arolas E, González-Ladrón-de-Guevara F. Towards an integrated crowdsourcing definition. J Inf Sci 2012 Mar 09;38(2):189-200. [doi: 10.1177/0165551512437638]
32. Cocos A, Qian T, Callison-Burch C, Masino AJ. Crowd control: Effectively utilizing unscreened crowd workers for biomedical data annotation. J Biomed Inform 2017 May;69:86-92 [FREE Full text] [doi: 10.1016/j.jbi.2017.04.003] [Medline: 28389234]
33. Wang C, Han L, Stein G, Day S, Bien-Gund C, Mathews A, et al. Crowdsourcing in health and medical research: a systematic review. Infect Dis Poverty 2020 Jan 20;9(1):8 [FREE Full text] [doi: 10.1186/s40249-020-0622-9] [Medline: 31959234]
34. Ørting SN, Doyle A, Van Hilten A, Hirth M, Inel O, Madan CR, et al. A survey of crowdsourcing in medical image analysis. Hum Comp 2020 Dec 01;7:1-26. [doi: 10.15346/hc.v7i1.1]
35. Deng J, Dong W, Socher R. ImageNet: A large-scale hierarchical image database. 2009 Presented at: IEEE Conference on Computer Vision and Pattern Recognition; Jun 20-25, 2009; Miami, FL. [doi: 10.1109/cvpr.2009.5206848]
36. Feng S, Woo MJ, Kim H. A game-based crowdsourcing platform for rapidly training middle and high school students to perform biomedical image analysis. In: Proceedings Volume 9699, Optics and Biophotonics in Low-Resource Settings II. 2016 Presented at: SPIE BiOS; Feb 13-18, 2016; San Francisco, CA p. 2016. [doi: 10.1117/12.2212310]
37. Morschheuser B, Hamari J, Koivisto J, Maedche A. Gamified crowdsourcing: Conceptualization, literature review, and future agenda. Int J Hum Comput Stud 2017 Oct;106:26-43. [doi: 10.1016/j.ijhcs.2017.04.005]
38. Bittel S, Roethlingshoefer V, Kenngott H, Wagner M, Bodenstedt S, Speidel S, et al. How to create the largest in-vivo endoscopic dataset. In: Intravascular Imaging and Computer Assisted Stenting, and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis. 2017 Presented at: 6th Joint International Workshops, CVII-STENT 2017 and Second International Workshop, LABELS 2017, Held in Conjunction with MICCAI; Sep 10–14, 2017; Quebec City, QC.
39. King AJ, Gehl RW, Grossman D, Jensen JD. Skin self-examinations and visual identification of atypical nevi: comparing individual and crowdsourcing approaches. Cancer Epidemiol 2013 Dec;37(6):979-984 [FREE Full text] [doi: 10.1016/j.canep.2013.09.004] [Medline: 24075797]
40. Tkaczyk ER, Coco JR, Wang J, Chen F, Ye C, Jagasia MH, et al. Crowdsourcing to delineate skin affected by chronic graft-vs-host disease. Skin Res Technol 2019 Jul;25(4):572-577 [FREE Full text] [doi: 10.1111/srt.12688] [Medline: 30786065]
41. Marghoob A, Braun R. Atlas of Dermoscopy, 2nd Ed. London, UK: CRC Press; 2012.
42. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012;22(3):276-282 [FREE Full text] [Medline: 23092060]
43. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977 Mar;33(1):159-174. [doi: 10.2307/2529310] [Medline: 843571]
44. DiagnosUs. Apple App Store. URL: https://apps.apple.com/us/app/diagnosus/id1369759559 [accessed 2022-12-19]
45. Kiani R, Corthell L, Shadlen M. Choice certainty is informed by both evidence and decision time. Neuron 2014 Dec 17;84(6):1329-1342 [FREE Full text] [doi: 10.1016/j.neuron.2014.12.015] [Medline: 25521381]
46. Argenziano G, Soyer HP, Chimenti S, Talamini R, Corona R, Sera F, et al. Dermoscopy of pigmented skin lesions: results of a consensus meeting via the Internet. J Am Acad Dermatol 2003 May;48(5):679-693. [doi: 10.1067/mjd.2003.281] [Medline: 12734496]

47. Patel P, Khanna S, McLellan B, Krishnamurthy K. The need for improved dermoscopy training in residency: a survey of US dermatology residents and program directors. Dermatol Pract Concept 2017 Apr;7(2):17-22 [FREE Full text] [doi: 10.5826/dpc.0702a03] [Medline: 28515987]
48. Soyer HP, Kenet RO, Wolf IH, Kenet BJ, Cerroni L. Clinicopathological correlation of pigmented skin lesions using dermoscopy. Eur J Dermatol 2000;10(1):22-28. [Medline: 10694293]
49. Braun R, Gaide O, Oliviero M, Kopf A, French L, Saurat J, et al. The significance of multiple blue-grey dots (granularity) for the dermoscopic diagnosis of melanoma. Br J Dermatol 2007 Nov;157(5):907-913. [doi: 10.1111/j.1365-2133.2007.08145.x] [Medline: 17725673]
50. Argenziano G, Zalaudek I, Corona R, Sera F, Cicale L, Petrillo G, et al. Vascular structures in skin tumors: a dermoscopy study. Arch Dermatol 2004 Dec;140(12):1485-1489. [doi: 10.1001/archderm.140.12.1485] [Medline: 15611426]
51. McKenna MT, Wang S, Nguyen TB, Burns JE, Petrick N, Summers RM. Strategies for improved interpretation of computer-aided detections for CT colonography utilizing distributed human intelligence. Med Image Anal 2012 Aug;16(6):1280-1292 [FREE Full text] [doi: 10.1016/j.media.2012.04.007] [Medline: 22705287]
52. Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Phys Ther 2005 Mar;85(3):257-268 [FREE Full text] [doi: 10.1093/ptj/85.3.257] [Medline: 15733050]
53. Candido Dos Reis FJ, Lynn S, Ali HR, Eccles D, Hanby A, Provenzano E, et al. Crowdsourcing the general public for large scale molecular pathology studies in cancer. EBioMedicine 2015 Jul;2(7):681-689 [FREE Full text] [doi: 10.1016/j.ebiom.2015.05.009] [Medline: 26288840]
54. Albarqouni S, Baur C, Achilles F, Belagiannis V, Demirci S, Navab N. AggNet: deep learning from crowds for mitosis detection in breast cancer histology images. IEEE Trans Med Imaging 2016 May;35(5):1313-1321. [doi: 10.1109/TMI.2016.2528120] [Medline: 26891484]
55. Cabitza F, Campagner A, Albano D, Aliprandi A, Bruno A, Chianca V, et al. The elephant in the machine: proposing a new metric of data reliability and its application to a medical case to assess classification reliability. Applied Sciences 2020 Jun 10;10(11):4014 [FREE Full text] [doi: 10.3390/app10114014]
56. Fujisawa Y, Inoue S, Nakamura Y. The possibility of deep learning-based, computer-aided skin tumor classifiers. Front Med (Lausanne) 2019;6:191 [FREE Full text] [doi: 10.3389/fmed.2019.00191] [Medline: 31508420]
57. Maier-Hein L. Crowd-algorithm collaboration for large-scale endoscopic image annotation with confidence. In: Ourselin S, Joskowicz L, Sabuncu M, Unal G, Wells W, editors. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016. MICCAI 2016. Lecture Notes in Computer Science(), vol 9901. Cham, Switzerland: Springer; 2016.

Abbreviations

CT: computed tomography
EASY: Expert Agreement on the Presence and Spatial Location of Melanocytic Features in Dermoscopy
IRR: interrater reliability
ISIC: International Skin Imaging Collaboration
ML: machine learning

Edited by C Lovis; submitted 04.04.22; peer-reviewed by Z Li, W Yu Jen , W Van Stoecker; comments to author 15.08.22; revised version received 28.09.22; accepted 16.10.22; published 18.01.23.

Please cite as:
Kentley J, Weber J, Liopyris K, Braun RP, Marghoob AA, Quigley EA, Nelson K, Prentice K, Duhaime E, Halpern AC, Rotemberg V Agreement Between Experts and an Untrained Crowd for Identifying Dermoscopic Features Using a Gamified App: Reader Feasibility Study
JMIR Med Inform 2023;11:e38412
URL: https://medinform.jmir.org/2023/1/e38412
doi:10.2196/38412
PMID:36652282

Original Paper

Toward Individualized Prediction of Binge-Eating Episodes Based on Ecological Momentary Assessment Data: Item Development and Pilot Study in Patients With Bulimia Nervosa and Binge-Eating Disorder

Ann-Kathrin Arend[1]; Tim Kaiser[2]; Björn Pannicke[1]; Julia Reichenberger[1]; Silke Naab[3]; Ulrich Voderholzer[3,4,5]; Jens Blechert[1]

[1]Department of Psychology, Centre for Cognitive Neuroscience, University of Salzburg, Salzburg, Austria

[2]Department of Clinical Psychology, University of Greifswald, Greifswald, Germany

[3]Schoen Clinic Roseneck, Prien am Chiemsee, Germany

[4]Department of Psychiatry and Psychotherapy, University Hospital, Ludwig Maximilian University of Munich, Munich, Germany

[5]Department of Psychiatry and Psychotherapy, University Hospital of Freiburg, Freiburg, Germany

Corresponding Author:
Ann-Kathrin Arend
Department of Psychology
Centre for Cognitive Neuroscience
University of Salzburg
Hellbrunnerstraße 34
Salzburg, 5020
Austria
Phone: 43 66280445102
Fax: 43 66280445126
Email: ann-kathrin.arend@plus.ac.at

Abstract

Background: Prevention of binge eating through just-in-time mobile interventions requires the prediction of respective high-risk times, for example, through preceding affective states or associated contexts. However, these factors and states are highly idiographic; thus, prediction models based on averages across individuals often fail.

Objective: We developed an idiographic, within-individual binge-eating prediction approach based on ecological momentary assessment (EMA) data.

Methods: We first derived a novel EMA-item set that covers a broad set of potential idiographic binge-eating antecedents from literature and an eating disorder focus group (n=11). The final EMA-item set (6 prompts per day for 14 days) was assessed in female patients with bulimia nervosa or binge-eating disorder. We used a correlation-based machine learning approach (Best Items Scale that is Cross-validated, Unit-weighted, Informative, and Transparent) to select parsimonious, idiographic item subsets and predict binge-eating occurrence from EMA data (32 items assessing antecedent contextual and affective states and 12 time-derived predictors).

Results: On average 67.3 (SD 13.4; range 43-84) EMA observations were analyzed within participants (n=13). The derived item subsets predicted binge-eating episodes with high accuracy on average (mean area under the curve 0.80, SD 0.15; mean 95% CI 0.63-0.95; mean specificity 0.87, SD 0.08; mean sensitivity 0.79, SD 0.19; mean maximum reliability of r_D 0.40, SD 0.13; and mean r_{CV} 0.13, SD 0.31). Across patients, highly heterogeneous predictor sets of varying sizes (mean 7.31, SD 1.49; range 5-9 predictors) were chosen for the respective best prediction models.

Conclusions: Predicting binge-eating episodes from psychological and contextual states seems feasible and accurate, but the predictor sets are highly idiographic. This has practical implications for mobile health and just-in-time adaptive interventions. Furthermore, current theories around binge eating need to account for this high between-person variability and broaden the scope of potential antecedent factors. Ultimately, a radical shift from purely nomothetic models to idiographic prediction models and theories is required.

(JMIR Med Inform 2023;11:e41513) doi:10.2196/41513

KEYWORDS

idiographic; individualized; N of 1; Ecological Momentary Assessment (EMA); Just-In-Time Adaptive Intervention (JITAI); binge eating; literature research; focus group; prediction algorithm; machine learning; Best Items Scales that are Cross-validated, Unit-weighted, Informative and Transparent; BISCUIT

Introduction

Binge Eating

Binge eating (objectively excessive food intake accompanied by feelings of loss of control) represents a core symptom of bulimia nervosa (BN), binge-eating disorder (BED), and the binge-purge subtype of anorexia nervosa. It is also the most debilitating symptom in most eating disorders (alongside the associated compensatory behavior in BN and binge-purge subtype of anorexia nervosa), accounting for gastrointestinal comorbidities, along with psychological consequences (eg, shame, secrecy, and social isolation [1]). Thus, interventions have focused on binge eating to ameliorate psychological consequences and subsequent purging behavior, which further contributes to oral and dental harms. However, treatment as usual–cognitive behavioral therapy for eating disorders (EDs) is effective for only about 65% of individuals with an ED [2] and has high relapse rates (26.8% across EDs) [3].

Nomothetic Binge-Eating Models

To predict binge eating, researchers typically rely on *nomothetic theories*—theories that are based on the average characteristics of multiple individuals in groups. Some nomothetic findings hold that individuals with BN and BED overeat in response to negative emotions, whereas healthy controls do not [4]. However, although particular efforts have been directed at predicting high-risk states for binge eating based on a variety of measures (eg, negative emotions or irregular eating patterns) [5,6], nomothetic binge-eating models often fail to translate to an idiographic–*individual*–level [7-10]. To illustrate, nomothetic theories claiming that emotional eating underlies binge eating [11] imply that emotion regulation interventions provide causative help [12,13]. However, this reasoning might not be applicable to patients who are prone to binge eating when impulsive, after extensive fasting periods, or experiencing dissociative states [6,13,14]. Correspondingly, various nomothetic theories of binge eating have proliferated. They differ substantially in the assumed causal mechanisms, which include, but are not limited to, emotional eating, impulsivity, restrained eating, food addiction, ego depletion, associative learning, and emotion-regulation or coping with emotions [4,15-21].

Idiographic Binge-Eating Models and Interventions

As binge eating can be highly impulsive, automatic, and difficult to resist, interventions that target binge eating based on its antecedents are promising as they attempt to stop the process as soon as possible, before the binge-eating pressure builds up. As such states can fluctuate quickly, they need to be assessed and evaluated with a high timely resolution to inform about the appropriate timing for interventions for high-risk states. Recently, the methodologies of just-in-time adaptive interventions (JITAIs) [22] and high-frequency ecological momentary assessment (EMA) have merged into a methodological framework that can be applied to binge-eating prediction and prevention. JITAIs have been shown to enhance cognitive behavioral therapy in BED and BN [23] and have been successfully implemented in other domains of eating behavior (eg, in weight loss) [24].

In the "OnTrack" weight-loss intervention, Forman et al [24] sampled emotions and stress, next to eating history and context conditions such as watching television or alcohol consumption. By investigating a wide range of antecedents for dietary lapses, they go well beyond what emotional-eating theory suggests as predictors (eg, negative emotions). Similarly, with their "Think Slim" app, Spanakis et al [25] showed that a sample of participants with normal weight and overweight can be clustered into multiple groups according to the different momentary states in which they tend to eat unhealthily. Therefore, applying a single nomothetic binge-eating theory might be insufficient to identify a broad spectrum of individually varying antecedents and would yield inaccurate predictions of binge eating in most individuals [26]. Instead, to cover all relevant antecedents for many patients, a broad set of EMA items is required. Notably, items that serve this purpose in weight disorders (eg, "OnTrack" JITAI-enhanced weight-loss intervention by Forman et al [24]) may not cover all antecedent states that arise in patients with clinical binge eating. Furthermore, despite being often disregarded within nomothetic frameworks, protective factors (eg, positive emotions or healthy coping [27,28]) have the potential to improve prediction accuracies in idiographic machine learning frameworks because of their negative associations with binge-eating likelihood. However, to balance participant burden with broad sampling, a baseline phase with the full item set could be followed by a phase with a reduced EMA-item set, based on a prediction model that identifies the ideographic subset of items that best predict binge eating for a given individual.

Aims and Hypothesis

This study examined the feasibility of the first part of this approach, that is, whether subsets of items could be found with good prediction accuracies for binge eating.

Furthermore, 2 studies were conducted to establish a conceptual and empirical foundation for JITAIs on binge eating. In study 1, we collected a comprehensive set of binge-eating antecedents in the form of EMA items. We combined a literature review with qualitative and quantitative interviews (focus group with 11 inpatients) following Soyster and Fisher [29]. In study 2, an algorithm was used to select idiographic subsets of binge-eating predictors based on Elleman et al [30], Kaiser et al [10], and Soyster et al [31]. We hypothesized that these idiographic binge-eating antecedents would predict binge eating with high

accuracy. This selection and prediction were tested in 13 patients with BN or BED.

Methods

Ethics Approval

All participants signed an informed consent form (stating which data were stored, where and for how long, who the investigator was, and the purpose of the study) approved by the ethics committee of the University of Salzburg (EK-GZ: 37/2018).

Study 1—Development of the EMA-Item Set

Literature Research

A PhD-level researcher systematically searched Google Scholar, PsycINFO, and PubMed databases for articles with the word "binge" in their title and the terms "ecological momentary" or "experience sampling" to find risk state descriptors with relevance to binge eating in the literature. The search resulted in 509 articles that were deduplicated and scanned for relevance. Only empirical articles reporting the results of EMA studies on binge eating were retained. A total of 262 articles were subsequently analyzed (see Multimedia Appendix 1, Figure S1 for an attrition diagram).

Text Analysis Using Word Embedding

Abstracts of all articles in the literature were retrieved. The R package *text2vec* [32] was used to perform global vector word-embedding analysis on these abstracts. Word embedding is an "unsupervised" learning algorithm that maps words to a vector space based on their similarity. It is unsupervised as no labeling of training data is needed because training is performed on aggregated global word-word cooccurrence statistics from a corpus. A matrix is calculated where each element X_{ij} represents how often word$_i$ appears in the context of word$_j$ (ie, in the same sentence). Thus, words can be represented numerically and their similarities can be compared [32].

The following parameters were set for training the word vectors (vector dimensions=100, window size=15, and minimum word count to be included in the model=5). The English stop words were removed. Single words (eg, "sadness"), as well as combinations of 2 words (eg, "negative affect"), were allowed in the model. The cosine similarity between word vectors was used to quantify the similarity between word embeddings. This metric computes the angle between 2 vectors to quantify the similarity in the vector space they inhabit. The interpretation of cosine similarity resembles that of the correlation coefficients. Perfectly similar word vectors have a cosine similarity of 1, whereas perfectly dissimilar vectors have a similarity of −1. We calculated the cosine similarity of all retained words with the words "binge" or "binges" retaining only words that had at least a cosine similarity of +.10 or −.10 (resembling a small effect according to the criteria of Cohen [33] for the interpretation of correlation coefficients). In this way, we intended to find words that were conceptually similar to "binge eating" while covering a wide range of binge-eating antecedents.

Integration Into a Preliminary Item List

In the next step, 2 authors independently rated whether a given retained word was quantifiable with a psychometric item (ie, the words "dissociation" or "dissociative" were rated as quantifiable with the item "I feel detached from myself." [0=*not at all* to *very much*=100]) and in terms of usefulness for an EMA survey. Items were only retained if they were rated as quantifiable and useful by both authors. Overlapping constructs were organized into categories to reduce redundancy. Finally, a preliminary list of 47 items was compiled from the empirical and theoretical constructs and complemented by constructs derived from previous EMA studies (Multimedia Appendix 2, Table S1).

Patient Focus Group

A focus group of inpatients (11 female adolescents and young adults in treatment for regular binge-eating episodes at the Schoen Clinic Roseneck, Germany) complemented this literature-based approach. It was conducted to tap into antecedents that nomothetic EMA research might have overlooked so far. After an individual written brainstorming session on "triggers and circumstances associated with binge eating," the inpatients rated the preliminary list of EMA items on relevance to their binge-eating episodes ("happens before/during/after binge eating…": 1=*[almost] never*, 3=*might or might not*, 5=*[almost] always*). A moderated discussion of the brainstormed and provided items concluded the sessions.

Next, 2 researchers analyzed the rating data and integrated patient-generated items. This led to the following changes: several constructs missing in the preliminary item list were identified and items were added to cover these gaps (eg, eating based on internal opposed to external motivation: "Did you eat on your own accord?"; (not) following a regular meal structure: "How much did you follow a regular meal structure today?"; and restricting specific foods: "Are you restricting on certain foods right now?").

The focus group participants further rated 27 of the provided items as positively associated with their binge-eating episodes (mean >3.5), 11 items as negatively associated (mean <2.5), and 9 items as unrelated to their binge-eating episodes (mean 2.5-3.5; Multimedia Appendix 2, Table S1). Some items were scored as unrelated (eg, "Right now I feel: tired" and "I engaged in increased levels of sport."), and items with large SDs (SD >1.00; eg, "Right now I feel: relived," "Right now I am shopping for groceries." and "I acted upon my plans regarding my eating behavior.") were disregarded, merged (eg, "I am in company." with "I am on my own."), or exchanged (eg, "I feel strained due to...work / university / school; close social network; wider social network; everyday stressors" with "Do you feel like you can handle all upcoming tasks and problems?"). As the patients expressed concerns over the redundancy of emotional states, 4 more items were disregarded ("Right now I feel: calm/ashamed/guilty/frustrated"). Finally, 4 items regarding eating behaviors such as "resistance to food craving" or "restriction" were rephrased to map more accurately on constructs introduced by the focus group (see Multimedia Appendix 3, Figure S1 for all item iterations)

Feedback of Clinicians

Finally, clinicians with experience in ED treatment (n=4) provided feedback on the gaps in the included constructs. This feedback was integrated by adding concepts such as accessibility to tasty food, day structure (ie, regular sleep and eating patterns), self-regulation intentions, and eating alone. This feedback further led us to include the autoregressive effect of binge-eating episodes on subsequent binge-eating risk in our models [34].

First Pilot

The EMA items were then piloted by 2 authors and 1 female patient with BN (consistent with the Diagnostic Statistical Manual-5 [DSM-5] [1]) to evaluate content, coverage, wording, and participant burden. Piloting revealed that some items needed further changes to map more accurately on the intended constructs: 1 item about adaptive coping strategies was added ("How much did you try to distract yourself from a possible urge to overeat by *healthy* strategies [e.g., relaxation, social activity, mindfulness, etc.]?") to complement the items on dysfunctional coping and distraction strategies, which were merged into one item ("how much did you try to distract yourself from a possible urge to overeat by unhealthy strategies [eg, alcohol, cigarettes, drugs, self-harm, etc]?"). Two items were rephrased, and 1 item assessing food craving was split up and rephrased to differentiate *food craving*, *overeating*, and *objective binge-eating episodes* (food craving: "how strong is your craving for certain foods right now?"; overeating: "how strong is your urge to overeat right now?"; and binge-eating episodes: "how high would you rate your risk for a binge-eating episode right now?").

The highly compliant participant with BN (all 84 EMA signals answered) reported that the participant burden was too high. Thus, 6 more items were disregarded to shorten the extensive list of items assessing different forms of self-licensing [35,36] and restrictions. Finally, the authors integrated the information gathered in the previous steps (ie, literature review, feedback of the focus group, feedback from clinicians, and feedback of the pilot patient) to make final iterations to the EMA-item set (see Multimedia Appendix 3, Figure S1 for all item iterations, and Multimedia Appendix 4, Tables S1 and S2 for the final EMA-item set).

Study 2—Idiographic Predictor Selection and Prediction of Binge Eating From EMA Data

Participants

Female patients with current BN (n=12) or BED (n=1) were recruited via mail from the waiting list for inpatient treatment of the Schoen Clinic Roseneck, Germany (n=10), and from web-based forums on eating disorders and psychology (n=3; see Multimedia Appendix 5, Figure S1 for a CONSORT [Consolidated Standards of Reporting Trials] flowchart). This study was advertised as a pilot study for a smartphone-based binge-eating intervention. The data were collected between April 2020 and April 2021.

Procedure

All participants completed the following study protocol. First, the BN and BED research diagnoses according to DSM-5 [1] were determined via telephone using the Eating Disorder Examination interview [37] and the Structured Clinical Interview for DSM-IV [38]. Both interviews were adapted to the diagnostic criteria of the DSM-5 (eg, 1 binge-eating episode per week for 3 months instead of 2 binges per week for 3 months).

The participants were then introduced to the EMA items and logged into the customized smartphone app *SmartEater*. SmartEater was used during the subsequent EMA phase, in which signal-based EMA questionnaires were inquired up to 84 times per participant (6 signal-contingent prompts per day, in intervals of 2.5 hours for 2 weeks; questionnaires expired 1 hour after the initial prompt). In addition, an event-contingent EMA questionnaire on overeating, loss of control, and binge-eating episodes was accessible. Participants were instructed to fill in this event-contingent questionnaire whenever they felt like they overate or felt a sense of loss of control over food intake or both. The event-contingent questionnaire included questions to differentiate between subjective and objective binge eating and objective overeating (Multimedia Appendix 4, Table S2). EMA items assessing emotions were presented in a randomized order. However, the other items were presented in a fixed order to prevent carryover effects. The participants were able to review and change their answers through a "back" button. Answering all items (except branched items) was mandatory for submission of the questionnaires.

After the EMA phase of 2 weeks, a JITAI phase of 2 weeks started, in which the participants received short intervention suggestions from the app to prevent binge-eating episodes at ideographically predicted high-risk times. Every study stage was accompanied by web-based questionnaires that assessed current eating behavior pathology, demographic data, perceived acceptability, feasibility, and so on. Data from the intervention phase were not covered in the present article. For reimbursement, the participants received €30 (US $32.80) and personalized feedback on their EMA data and psychometric web-based questionnaires.

Data Preparation and Measures

To avoid the violation of the assumption of equally spaced time series [39], empty rows were inserted in the data set after every last signal for a given day. This prevented the prediction algorithm from regressing on data from the previous day.

Binge-Eating Episodes–Criterion

Objective binge-eating episodes, characterized by (1) "feelings of loss of control over eating behavior" and (2) "consumption of objectively large, inappropriate amounts of food" [1,37,40], were identified from eating episodes reported over the signal-based (1 item: "Was your meal a main meal**,** snack, or binge?") and event-based EMA questionnaires (2 items: "Would other people rate the amount of food as excessive under similar circumstances?" and "Did you feel like you are losing control of your eating behavior?"). The signal-based and the 2 event-based items were recoded into a binary variable indicating the occurrence of an objective binge-eating episode (binge-eating episode reported=1, no binge-eating episode reported=0). As the algorithm was supposed to predict future

binge-eating episodes, this variable was shifted backward in time by one signal (approximately 2.5 hour).

EMA Predictors

An unshifted version of the binge-eating variable was included as a possible predictor of the autoregressive effects of binge eating. Furthermore, additional EMA items (n=31) were used to model possible binge-eating antecedents. Thus, only items that were assessed with every signal-based questionnaire were included (aside from the binge-eating classifier), as each variable needed to have a sufficient percentage of data points within a person (see Multimedia Appendix 4, Table S1 for the wording of each item).

Time Predictors

Time variables, especially in the form of circles and distinct times of day, have been shown to be highly predictive in everyday life [41]. EMA studies have even found peak times for certain binge-eating antecedents (ie, food cravings or hunger; [42]) and binge eating itself [43-45]. Thus, as temporal data are passively collected in the EMA setting via timestamps, without additional participant input, we decided to include different temporal predictors that could detect a single high-risk time per day (24-hour oscillation) or several times per day (sub-24 hour oscillation).

Variables representing 8-, 12- and 24-hour sinusoidal and cosinusoidal cycles were computed based on the cumulative sum of time differences between assessments (eg, 10:30 AM-8 AM, 1 PM-10:30 AM=2.5, 5, 7.5...). For example, a 24-hour sinusoid cycle was calculated using the following formula: $sin24h = sin(2\pi : 24 * \Delta_t)$, where Δ_t is the difference between assessment points in hours (here: 2.5). Finally, dummy-coded variables representing the time of day were calculated for each signal (morning, late morning, early afternoon, afternoon, evening, and late evening). This allows for identifying a daytime when binge eating is particularly likely for a given participant (eg, when returning from work) that could not be well captured by the cyclical predictors.

Application of the Best Items Scale that is Cross-validated, Unit-weighted, Informative, and Transparent Algorithm to EMA Data (for Idiographic Predictor Selection and Prediction of Binge Eating)

The machine learning algorithm Best Item Scales that are Cross-validated, Unit-weighted, Informative and Transparent (BISCUIT) [30] of the *bestScales* function from the R package *psych* [46] was applied separately to the EMA data of each patient to select the best idiographic predictors of binge-eating episodes. This method was chosen because of its (1) robustness to missing data; (2) use of unit-weighted scoring of predictors, which was found to be more generalizable, especially in the context of prediction; and (3) tendency to select more parsimonious predictor sets compared with other approaches such as Elastic Net regression [30,47,48]. BISCUIT is a simple algorithm that correlates a set of predictors (here all EMA and time variables) with a criterion (here, the binary time-shifted binge-eating variable at t+1) and retains the predictors with the highest correlation to form a unit-weighted scale [10,30,46]. This scale was then used to estimate the out-of-sample predictive performance using 10-fold cross-validation. The average correlation of the scale with the criterion across 10 cross-validation splits was then computed, and the set of items with the highest cross-validated correlation was retained [30,46]. The output of BISCUIT is the selection of items showing maximum predictive validity, as the cutoff values that lead to the highest combination of sensitivity and specificity are retained [30,46].

Thus, multiple Rs (pairwise Pearson correlations) of all predictors with time-shifted binge-eating episodes (at t+1) were calculated for each participant separately to select the idiographic predictor sets. Furthermore, the area under the curve (AUC) with a bootstrapped 95% CI, specificity, sensitivity, and within- and out-of-sample reliability were calculated as prediction accuracy measures of the idiographic predictor sets and their prediction of binge eating in the next 2.5 hours (t+1).

Results

Study 1—EMA-Item Set

The final signal-contingent EMA questionnaire included 36 EMA items (momentary emotions, stress, exhaustion, and context; eg, being alone, social interactions, dissociations, eating behavior, resistance to food craving, distraction, and coping), which were designed to be assessed 6 times per day. In addition, an optional event-contingent EMA questionnaire on overeating, loss-of-control eating, and binge eating was self-initialized and included 20 items. See Multimedia Appendix 4, Tables S1 and S2 for all interval- and event-contingent items and their wording. A flowchart of all iterations applied to the EMA-item set can be found in Multimedia Appendix 3, Figure S1).

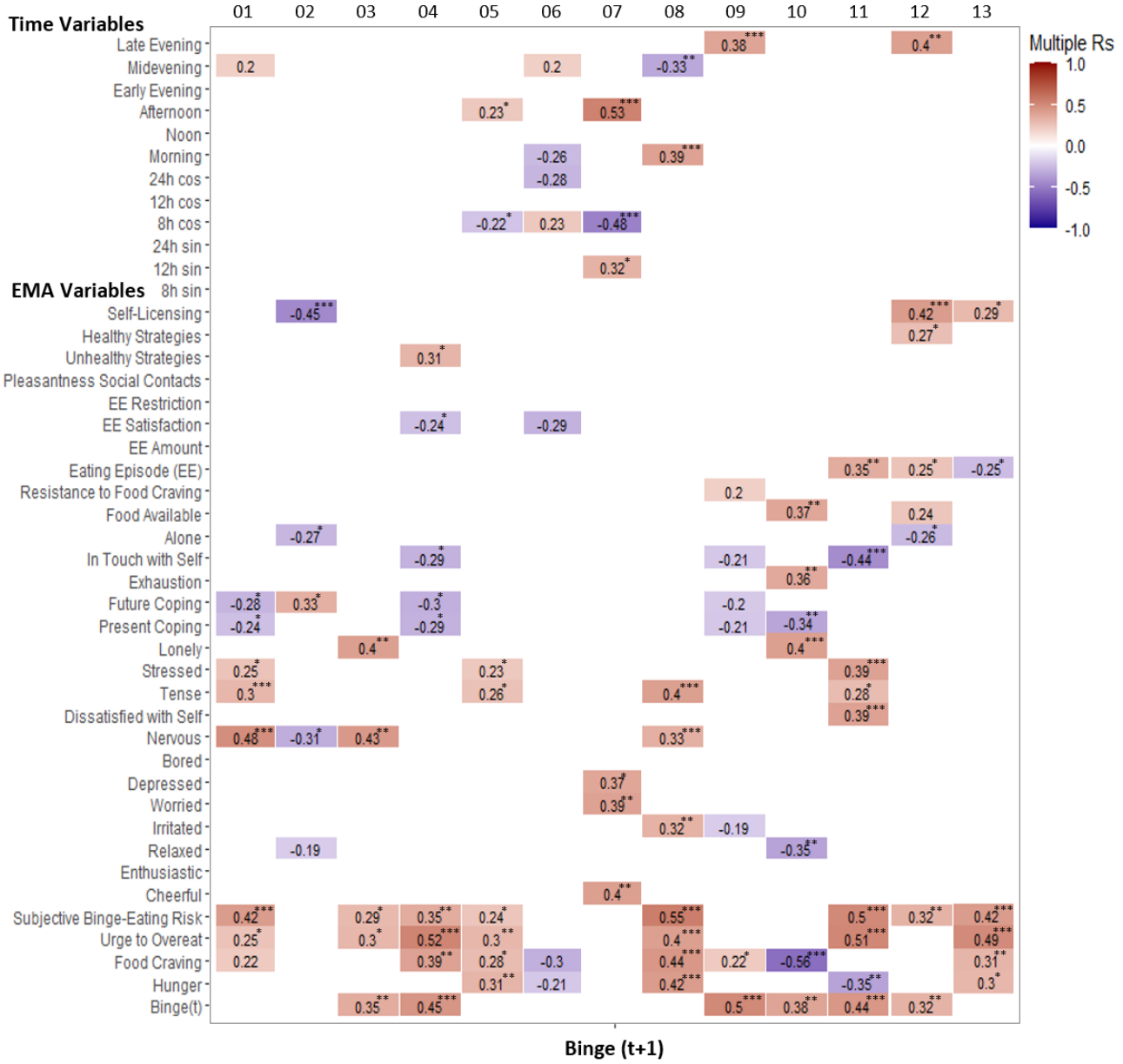
Study 2—Idiographic Predictor Selection and Prediction of Binge Eating From EMA Data (by Application of the BISCUIT Algorithm)

Selection of Idiographic Predictor Subsets

The patients (n=13) answered on an average 67.3 out of 84 EMA prompts (SD 13.4; range 43-84; see Multimedia Appendix 6, Table S1 for EMA compliance and occurrences of binge-eating episodes per patient). Across participants, the algorithm selected highly heterogeneous predictor sets of varying sizes (mean 7.31, SD 1.49; range 5-9 predictors) for the prediction of binge-eating episodes.

Figure 1 shows the idiographic predictor selection that showed maximum predictive validity for each participant. Thus, the predictors (at t) with the highest multiple Rs (pairwise Pearson correlations) with time-shifted binge-eating episodes (at t+1) were selected. All listed items were selected as idiographic predictors of binge eating, independent of their significance. However, we additionally calculated the significance of the correlations for the context. The exact *P* values, codes, and data can be found in the corresponding project in the Open Science Framework [49]. Note that the results might vary slightly, as the R function *set.seed* does not apply to the cross tables.

Figure 1. Idiographic predictor subsets for binge eating with Pairwise Pearson Correlations (Multiple Rs) of each selected predictor of binge eating in the next 2.5 hours (t+1). *, **, and *** indicate that the correlations are significant at a level of .05, .01, and .001, respectively; 2-tailed. EMA: ecologic momentary assessment.



Prediction of Binge Eating by Idiographic Predictor Subsets

The selection of idiographic predictor sets resulted in good average prediction accuracy (mean AUC 0.80, SD 0.15; mean 95% CI 0.63-0.95; mean specificity 0.87, SD 0.08; mean sensitivity 0.79, SD 0.19; mean maximum reliability of r_D 0.40, SD 0.13; mean r_{CV} 0.13, SD 0.31). The mean AUC of 0.80 indicates that there is on average an 80% chance that the idiographic models predict binge and nonbinge episodes accurately. The mean specificity of 0.87 indicates that the idiographic models mistakenly classified 13 of 100 episodes as binge-eating episodes. The mean sensitivity of 0.79 indicates that the idiographic models mistakenly classified 21 out of 100 binge-eating episodes as nonbinge episodes. Table 1 shows the prediction accuracies of the idiographic predictor subsets for binge-eating episodes per participant. R code and data are available from the Open Science Framework [49].

Table 1. Model fit indices for prediction of binge eating in the next 2.5 hours from idiographic predictors, selected by the Best Items Scale that is Cross-validated, Unit-weighted, Informative, and Transparent (BISCUIT) algorithm, separately for each participant.

	Participants												
	01	02	03	04	05	06	07	08	09	10	11	12	13
Model fit indices													
AUC[a] (95% CI)[b]	0.92 (0.75-1.00)	0.97 (0.92-1.00)	0.84 (0.70-0.98)	0.51 (0.23-0.80)	0.73 (0.45-1.00)	0.93 (0.77-1.00)	0.93 (0.83-1.00)	0.85 (0.75-0.95)	0.63 (0.41-0.85)	0.75 (0.58-0.93)	0.72 (0.53-0.92)	0.60 (0.29-0.90)	0.98 (0.94-1.00)
Specificity	0.84	0.89	0.74	1.00	0.86	0.85	0.90	0.81	1.00	0.74	0.87	0.91	0.96
Sensitivity	1.00	1.00	.86	.45	.67	1.00	1.00	.80	.56	.73	.69	.56	1.00
Derivation step (within-sample performance)													
r_D[c,d] (SD)	0.48 (0.04)	0.42 (0.05)	0.53 (0.04)	0.33 (0.05)	0.34 (0.03)	0.47 (0.06)	0.53 (0.06)	0.51 (0.10)	0.10 (0.22)	0.46 (0.06)	0.50 (0.06)	0.18 (0.10)	0.32 (0.23)
Validation step (out of sample performance)													
r_{CV}[d,e] (SD)	0.41 (0.34)	0.10 (0.56)	–0.02 (0.72)	0.36 (0.38)	0.38 (0.62)	–0.36 (0.28)	0.29 (0.30)	0.54 (0.42)	–0.11 (0.46)	0.34 (0.27)	0.27 (0.64)	0.08 (0.78)	–0.56 (0.57)

[a]AUC: area under the curve.

[b]CI: bootstrapped 95% CI of the AUC.

[c]r_D: multiple *R* of the unit-weighted scale in the derivation step.

[d]*r*: pairwise Pearson correlation of the item with time-shifted binge-eating episodes (at t+1).

[e]r_{CV}: average cross-validated multiple *R* of the derived scale.

Discussion

Principal Findings

Study 1—EMA-Item Set

This study used a mixed methods approach to develop a conceptual and statistical basis for an idiographic JITAI for binge eating. The EMA-item development in study 1 followed a replicable procedure similar to Soyster and Fisher [29] while considering nomothetic theories on binge-eating antecedents (ie, emotional eating) and underwent several qualitative (literature research and focus group brainstorming and discussion) and quantitative (focus group ratings) iterations and piloting.

This resulted in a broad EMA-item set (Multimedia Appendix 4), including several constructs underrepresented in the nomothetic literature (eg, “I feel detached from myself.”, and “specific” restrictions “did you restrict yourself [eg, by eating less, avoiding certain foods]?” [12,50,51]). This approach also helped us shorten the extensive lists of emotional states (eg, “right now I feel...calm/relived/ashamed/guilty/frustrated.”) because within-person ratings for similar emotions were often identical, and concerns about redundancy were expressed during the moderated discussion. Furthermore, we did not only incorporate risk factors into the EMA-item set but also protective factors that could potentially decrease the likelihood of binge eating (ie, healthy coping strategies to keep oneself from binge eating or positive emotions [27,28]). The role of protective factors is often overlooked in nomothetic binge-eating theories but is crucial to idiographic binge-eating prediction and intervention models.

Study 2—Prediction Based on Idiographic Predictor Subsets

Regularly completing extensive EMA-item sets (such as the present one with 36 interval-contingent and 20 event-contingent items) becomes increasingly burdensome over prolonged study periods. Thus, we applied a machine learning algorithm to the EMA data of patients with BN and BED to select parsimonious idiographic subsets of EMA items. This data-driven selection optimizes the predictive power within participants and decreases potential researcher bias.

The idiographic item subsets predicted binge-eating episodes with a high average accuracy (mean AUC 0.80) across 13 patients. Notably, the sensitivity approached 100% (successful prediction of every reported binge) in several patients, without forfeiting much specificity (predicting no binge when none occurred). This is noteworthy as outcome frequency was not extremely high (mean 10.4, SD 7.4; range 2-28 binge-eating episodes; see also Multimedia Appendix 6, Table S1 “Number of binge-eating episodes and total data points per participant”).

Secondary Findings

Regarding the composition of the selected item sets, a high selection rate of items with high proximity to the binge-eating construct was evident (ie, hunger, food craving, urge to overeat, subjective binge-eating risk, and preceding binge-eating episodes). This suggests that some patients may accurately predict upcoming binge-eating episodes. This reveals a relatively high level of insight into the temporal evolution of the symptoms in some patients. Surprisingly, hunger and food craving were negatively correlated with binge eating in 3 patients. One could speculate that the negative correlations between hunger and binge eating in patients 06 and 11 point to disinhibition, that is,

because of the temporary abandonment of rigid diet rules after eating in the absence of hunger [52,53].

In addition to items with conceptual similarity to binge eating, emotional items were selected in 9 patients. This supports the relevance of emotional eating in binge-eating predictions [4,11,19]. However, the selected emotion sets were highly heterogeneous across the 9 patients. In fact, no single emotion item (or specific set of emotion items) was consistently selected across all patients. This speaks against a singular and generalizable emotional eating theory of binge eating. Similarly, because no other nonemotion–related predictor was consistently selected across all patients, our pilot data provide no evidence for other generalizable nomothetic theories of binge eating. Thus, several nomothetic theories are needed to explain present heterogeneity, which may in turn explain the multitude of competing nomothetic binge-eating theories. Clearly, nomothetic theories must model individual differences more explicitly to account for these findings. These findings also support the use of a broad EMA-item set that covers a large range of possible binge-eating antecedents in the context of idiographic prediction [4-6,14-21].

Interestingly, time-derived predictors were selected only in 7 patients. In 6 of these patients, discrete time predictors were chosen that were consistent with the literature on the timing of binge-eating peaks (ie, afternoon to late evening) [43-45]. Time cycles were only selected in 3 patients. This is surprising given the observation of cyclic symptoms (eg, in depression [41]). However, time-based predictors may be more powerful if EMA items with conceptual similarity to binge eating are omitted. In the case of binge eating, time cycles could represent a rising and falling urge to overeat (eg, due to prolonged restriction between meals) [41,54]. Discrete time variables could represent the time of the day where a patient usually binges (eg, due to contexts such as being alone at home every afternoon) [41]. Assessing such time-derived variables does not require user input and thus does not contribute to the participant burden. This makes them valuable for the predictions in the JITAI framework.

Limitations and Strengths

Compared with the high average within-sample performance (mean r_D 0.40), the average out-of-sample performance (mean r_{CV} 0.13) was lower, suggesting limited out-of-sample generalizability. This might be because of 10-fold cross-validation, which does not account for the serial correlation and potential nonstationarity of time-series data [55]. Future studies could resort to alternative time-series–specific techniques (ie, roll forward cross-validation and out-of-sample evaluation) that ensure that training data always precede test data. However, X-fold cross-validation has been shown to outperform these techniques [55]. Furthermore, the number of observations was limited (max 84 per participant), leading to relatively small splits in the 10-fold cross-validation. Thus, there was a high possibility of randomly drawn training sets that were unrepresentative of the data set. The results from the validation step might also vary slightly, as the R function *set.seed* does not apply to the cross tables.

Another general drawback of the BISCUIT algorithm is that nonlinear trends and interaction effects among predictors are not considered. In addition, when applied under optimal conditions (ie, big data sets and no missing data), gold standard machine learning approaches, such as random forests [56] and XGBoost [57] in combination with super learners [58], calibrate better to the data. However, for typical EMA data sets, the conditions are rarely optimal for these algorithms. Missing data and a limited number of observations are typical features of high-burden EMA sampling schemes. However, BISCUIT was created to handle these problematic properties. BISCUIT outperformed random forest and elastic net approaches in other studies with smaller idiographic data sets and more missing data (Beck et al [59]: mean 57.4, SD 16.3; range 40-109 EMA observations; present data: mean 67.3, SD 13.4; range 43-84 EMA observations).

Finally, the idiographic approach used in this study precludes mechanistic and theoretical inferences about binge eating. Generally, machine learning algorithms are silent about the underlying mechanisms; instead, they tailor models are as close as possible to the given data and conditions. Thus, the present results are highly specific, for example, to the used “prediction interval” of 2.5 hours between predictors and outcome. This could be problematic as it has been shown that emotions and eating can influence each other at different time intervals [60].

Implications and Future Directions

In addition to emphasizing the importance of a broad predictor set, the results have a direct implication for the JITAI and EMA methodology: participant burden in longer EMA sampling periods precludes the use of large EMA-item sets. Thus, such EMA studies might prune their large EMA-item sets after a “calibration period” by applying the described predictor-selection approach. Therefore, the participant burden is reduced, whereas accurate idiographic binge-eating predictions are retained. Such predictions can then be used to trigger JITAIs, as done by Forman et al [24,61] in a JITAI on dietary lapses.

Future studies may transfer the present work to a range of disordered and maladaptive eating behaviors (eg, purging behaviors or food restrictions) to develop low-threshold JITAIs. EMA-item–based prediction should be compared with predictions generated from passive data sources (ie, smartphone sensors, use data, and wearable data) that do not inflict much user burden [10,62-64]. In the long term, acceptance, dropout rates, and effectiveness of JITAI protocols on binge eating need to be tested in microrandomized trials [65] and classic randomized controlled trials against nonadaptive, non–real-time interventions before the ultimate recommendation as the gold standard.

Finally, feeding back personal binge-eating predictors can serve as a psychoeducational intervention and raise awareness of personal risk and protective factors. Such personal binge-eating predictors can also inform conventional face-to-face psychotherapy. Patients with a clear dominance of emotion-related predictors might profit from emotion-focused interventions [66] more than patients with a dominance of

impulsive or craving-related predictors, who might profit more from impulse control intervention [67].

Acknowledgments

This project received funding from the European Research Council under the European Union's Horizon 2020 Research and Innovation Program (ERCStG-2014 639445 NewEat) and the Austrian Science Fund (KLI 762). Publishing open access was supported by the Paris Lodron University of Salzburg and Austrian Science Fund. The funding bodies were not involved in the study design, assessment and analysis of data, interpretation of results, or writing of the manuscript.

Authors' Contributions

AKA, TK, BP, JR, and JB conceptualized the studies; AKA conducted the studies; AKA, TK, and BP analyzed the data; AKA, TK, and JB wrote the paper; and BP, JR, SN, and UV contributed to the interpretation of the studies and critically revised the work for important intellectual content. All authors have read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Attrition diagram.
[DOCX File , 70 KB - medinform_v11i1e41513_app1.docx]

Multimedia Appendix 2
First item list.
[DOCX File , 63 KB - medinform_v11i1e41513_app2.docx]

Multimedia Appendix 3
Iterations in the creation of the item set.
[DOCX File , 82 KB - medinform_v11i1e41513_app3.docx]

Multimedia Appendix 4
Final ecologic momentary assessment items.
[DOCX File , 72 KB - medinform_v11i1e41513_app4.docx]

Multimedia Appendix 5
CONSORT (Consolidated Standards of Reporting Trials) flow diagram.
[DOCX File , 83 KB - medinform_v11i1e41513_app5.docx]

Multimedia Appendix 6
Number of binge-eating episodes and ecologic momentary assessment observations per patient.
[DOCX File , 58 KB - medinform_v11i1e41513_app6.docx]

References

1. American Psychiatric Association. Diagnostisches und Statistisches Manual Psychischer Störungen DSM-5®. Göttingen, Germany: Hogrefe; 2015.
2. Fairburn CG, Bailey-Straebler S, Basden S, Doll HA, Jones R, Murphy R, et al. A transdiagnostic comparison of enhanced cognitive behaviour therapy (CBT-E) and interpersonal psychotherapy in the treatment of eating disorders. Behav Res Ther 2015 Jul;70:64-71 [FREE Full text] [doi: 10.1016/j.brat.2015.04.010] [Medline: 26000757]
3. Helverskov JL, Clausen L, Mors O, Frydenberg M, Thomsen PH, Rokkedal K. Trans-diagnostic outcome of eating disorders: a 30-month follow-up study of 629 patients. Eur Eat Disord Rev 2010;18(6):453-463. [doi: 10.1002/erv.1025] [Medline: 20593480]
4. Cardi V, Leppanen J, Treasure J. The effects of negative and positive mood induction on eating behaviour: a meta-analysis of laboratory studies in the healthy population and eating and weight disorders. Neurosci Biobehav Rev 2015 Oct;57:299-309. [doi: 10.1016/j.neubiorev.2015.08.011] [Medline: 26299807]
5. Haedt-Matt AA, Keel PK. Revisiting the affect regulation model of binge eating: a meta-analysis of studies using ecological momentary assessment. Psychol Bull 2011 Jul;137(4):660-681 [FREE Full text] [doi: 10.1037/a0023660] [Medline: 21574678]

6. Zunker C, Peterson CB, Crosby RD, Cao L, Engel SG, Mitchell JE, et al. Ecological momentary assessment of bulimia nervosa: does dietary restriction predict binge eating? Behav Res Ther 2011 Oct;49(10):714-717 [FREE Full text] [doi: 10.1016/j.brat.2011.06.006] [Medline: 21764036]
7. Adolf J, Schuurman NK, Borkenau P, Borsboom D, Dolan CV. Measurement invariance within and between individuals: a distinct problem in testing the equivalence of intra- and inter-individual model structures. Front Psychol 2014 Sep 19;5:883 [FREE Full text] [doi: 10.3389/fpsyg.2014.00883] [Medline: 25346701]
8. Beltz AM, Wright AG, Sprague BN, Molenaar PC. Bridging the nomothetic and idiographic approaches to the analysis of clinical data. Assessment 2016 Aug;23(4):447-458 [FREE Full text] [doi: 10.1177/1073191116648209] [Medline: 27165092]
9. Fisher AJ, Medaglia JD, Jeronimus BF. Lack of group-to-individual generalizability is a threat to human subjects research. Proc Natl Acad Sci U S A 2018 Jul 03;115(27):E6106-E6115 [FREE Full text] [doi: 10.1073/pnas.1711978115] [Medline: 29915059]
10. Kaiser T, Butter B, Arzt S, Pannicke B, Reichenberger J, Ginzinger S, et al. Time-lagged prediction of food craving with qualitative distinct predictor types: an application of BISCWIT. Front Digit Health 2021 Sep 20;3:694233. [doi: 10.3389/fdgth.2021.694233]
11. Reichenberger J, Schnepper R, Arend AK, Richard A, Voderholzer U, Naab S, et al. Emotional eating across different eating disorders and the role of body mass, restriction, and binge eating. Int J Eat Disord 2021 May;54(5):773-784 [FREE Full text] [doi: 10.1002/eat.23477] [Medline: 33656204]
12. Engelberg MJ, Steiger H, Gauvin L, Wonderlich SA. Binge antecedents in bulimic syndromes: an examination of dissociation and negative affect. Int J Eat Disord 2007 Sep;40(6):531-536. [doi: 10.1002/eat.20399] [Medline: 17573684]
13. Hilbert A, Tuschen-Caffier B. Maintenance of binge eating through negative mood: a naturalistic comparison of binge eating disorder and bulimia nervosa. Int J Eat Disord 2007 Sep;40(6):521-530. [doi: 10.1002/eat.20401] [Medline: 17573697]
14. McShane JM, Zirkel S. Dissociation in the binge-purge cycle of bulimia nervosa. J Trauma Dissociation 2008;9(4):463-479. [doi: 10.1080/15299730802225680] [Medline: 19042792]
15. Evers C, Dingemans A, Junghans AF, Boevé A. Feeling bad or feeling good, does emotion affect your consumption of food? A meta-analysis of the experimental evidence. Neurosci Biobehav Rev 2018 Sep;92:195-208. [doi: 10.1016/j.neubiorev.2018.05.028] [Medline: 29860103]
16. Schag K, Schönleber J, Teufel M, Zipfel S, Giel KE. Food-related impulsivity in obesity and binge eating disorder--a systematic review. Obes Rev 2013 Jun;14(6):477-495. [doi: 10.1111/obr.12017] [Medline: 23331770]
17. Gearhardt AN, White MA, Potenza MN. Binge eating disorder and food addiction. Curr Drug Abuse Rev 2011 Sep;4(3):201-207 [FREE Full text] [doi: 10.2174/1874473711104030201] [Medline: 21999695]
18. Loth KA, Goldschmidt AB, Wonderlich SA, Lavender JM, Neumark-Sztainer D, Vohs KD. Could the resource depletion model of self-control help the field to better understand momentary processes that lead to binge eating? Int J Eat Disord 2016 Nov;49(11):998-1001 [FREE Full text] [doi: 10.1002/eat.22641] [Medline: 27768820]
19. Dingemans A, Danner U, Parks M. Emotion regulation in binge eating disorder: a review. Nutrients 2017 Nov 22;9(11):1274 [FREE Full text] [doi: 10.3390/nu9111274] [Medline: 29165348]
20. Jansen A. A learning model of binge eating: cue reactivity and cue exposure. Behav Res Ther 1998 Mar;36(3):257-272. [doi: 10.1016/s0005-7967(98)00055-2] [Medline: 9642846]
21. Bongers P, Jansen A. Emotional eating and Pavlovian learning: evidence for conditioned appetitive responding to negative emotional states. Cogn Emot 2017 Feb;31(2):284-297. [doi: 10.1080/02699931.2015.1108903] [Medline: 26539994]
22. Smith KE, Juarascio A. From Ecological Momentary Assessment (EMA) to Ecological Momentary Intervention (EMI): past and future directions for ambulatory assessment and interventions in eating disorders. Curr Psychiatry Rep 2019 Jun 04;21(7):53. [doi: 10.1007/s11920-019-1046-8] [Medline: 31161276]
23. Juarascio AS, Parker MN, Lagacey MA, Godfrey KM. Just-in-time adaptive interventions: a novel approach for enhancing skill utilization and acquisition in cognitive behavioral therapy for eating disorders. Int J Eat Disord 2018 Aug;51(8):826-830 [FREE Full text] [doi: 10.1002/eat.22924] [Medline: 30051495]
24. Forman EM, Goldstein SP, Crochiere RJ, Butryn ML, Juarascio AS, Zhang F, et al. Randomized controlled trial of OnTrack, a just-in-time adaptive intervention designed to enhance weight loss. Transl Behav Med 2019 Nov 25;9(6):989-1001. [doi: 10.1093/tbm/ibz137] [Medline: 31602471]
25. Spanakis G, Weiss G, Boh B, Lemmens L, Roefs A. Machine learning techniques in eating behavior e-coaching. Pers Ubiquit Comput 2017 Jun 8;21(4):645-659. [doi: 10.1007/s00779-017-1022-4]
26. Levinson CA, Vanzhula I, Brosof LC. Longitudinal and personalized networks of eating disorder cognitions and behaviors: targets for precision intervention a proof of concept study. Int J Eat Disord 2018 Nov;51(11):1233-1243. [doi: 10.1002/eat.22952] [Medline: 30291641]
27. Cardi V, Leppanen J, Leslie M, Esposito M, Treasure J. The use of a positive mood induction video-clip to target eating behaviour in people with bulimia nervosa or binge eating disorder: an experimental study. Appetite 2019 Feb 01;133:400-404. [doi: 10.1016/j.appet.2018.12.001] [Medline: 30529607]
28. Kelly NR, Lydecker JA, Mazzeo SE. Positive cognitive coping strategies and binge eating in college women. Eat Behav 2012 Aug;13(3):289-292. [doi: 10.1016/j.eatbeh.2012.03.012] [Medline: 22664415]

29. Soyster PD, Fisher AJ. Involving stakeholders in the design of ecological momentary assessment research: an example from smoking cessation. PLoS One 2019 May 22;14(5):e0217150 [FREE Full text] [doi: 10.1371/journal.pone.0217150] [Medline: 31116777]
30. Elleman LG, McDougald SK, Condon DM, Revelle W. That takes the BISCUIT: predictive accuracy and parsimony of four statistical learning techniques in personality data, with data missingness conditions. Eur J Psychol Assess 2020 Nov;36(6):948-958. [doi: 10.1027/1015-5759/a000590]
31. Soyster PD, Ashlock L, Fisher AJ. Pooled and person-specific machine learning models for predicting future alcohol consumption, craving, and wanting to drink: a demonstration of parallel utility. Psychol Addict Behav 2022 May;36(3):296-306. [doi: 10.1037/adb0000666] [Medline: 35041441]
32. Selivanov D, Bickel M, Wang Q. text2vec: Modern Text Mining Framework for R. The Comprehensive R Archive Network. 2022 Nov 30. URL: https://CRAN.R-project.org/package=text2vec [accessed 2022-12-05]
33. Cohen J. Statistical Power Analysis for the Behavioral Sciences. 2nd edition. Hillsdale, NJ, USA: Lawrence Earlbaum Associates; 1988.
34. Lavender JM, Utzinger LM, Cao L, Wonderlich SA, Engel SG, Mitchell JE, et al. Reciprocal associations between negative affect, binge eating, and purging in the natural environment in women with bulimia nervosa. J Abnorm Psychol 2016 Apr;125(3):381-386 [FREE Full text] [doi: 10.1037/abn0000135] [Medline: 26692122]
35. Prinsen S, Evers C, Wijngaards L, van Vliet R, de Ridder D. Does self-licensing benefit self-regulation over time? An ecological momentary assessment study of food temptations. Pers Soc Psychol Bull 2018 Jun;44(6):914-927 [FREE Full text] [doi: 10.1177/0146167218754509] [Medline: 29383977]
36. Witt Huberts JC, Evers C, De Ridder DT. License to sin: self-licensing as a mechanism underlying hedonic consumption. Eur J Soc Psychol 2012 Jun;42(4):490-496. [doi: 10.1002/ejsp.861]
37. Hilbert A, Tuschen-Caffier B, Ohms M. Eating disorder examination: deutschsprachige version des strukturierten essstörungsinterviews. Diagnostica 2004 Apr;50(2):98-106. [doi: 10.1026/0012-1924.50.2.98]
38. Wittchen HU, Zaudig M, Fydrich T. SKID. Strukturiertes Klinisches Interview für DSM-IV. Achse I und II. Göttingen, Germany: Hogrefe; 1997.
39. Jebb AT, Tay L, Wang W, Huang Q. Time series analysis for psychological research: examining and forecasting change. Front Psychol 2015 Jun 9;6:727 [FREE Full text] [doi: 10.3389/fpsyg.2015.00727] [Medline: 26106341]
40. International Statistical Classification of Diseases and Related Health Problems (11th revision). World Health Organization. Geneva, Switzerland: World Health Organization; 2019. URL: https://www.who.int/standards/classifications/classification-of-diseases [accessed 2022-05-09]
41. Fisher AJ, Bosley HG. Identifying the presence and timing of discrete mood states prior to therapy. Behav Res Ther 2020 May;128:103596. [doi: 10.1016/j.brat.2020.103596] [Medline: 32135317]
42. Reichenberger J, Richard A, Smyth JM, Fischer D, Pollatos O, Blechert J. It's craving time: time of day effects on momentary hunger and food craving in daily life. Nutrition 2018 Nov;55-56:15-20. [doi: 10.1016/j.nut.2018.03.048] [Medline: 29960151]
43. Smyth JM, Wonderlich SA, Sliwinski MJ, Crosby RD, Engel SG, Mitchell JE, et al. Ecological momentary assessment of affect, stress, and binge-purge behaviors: day of week and time of day effects in the natural environment. Int J Eat Disord 2009 Jul;42(5):429-436 [FREE Full text] [doi: 10.1002/eat.20623] [Medline: 19115371]
44. Stein RI, Kenardy J, Wiseman CV, Dounchis JZ, Arnow BA, Wilfley DE. What's driving the binge in binge eating disorder?: a prospective examination of precursors and consequences. Int J Eat Disord 2007 Apr;40(3):195-203. [doi: 10.1002/eat.20352] [Medline: 17103418]
45. Schreiber-Gregory DN, Lavender JM, Engel SG, Wonderlich SA, Crosby RD, Peterson CB, et al. Examining duration of binge eating episodes in binge eating disorder. Int J Eat Disord 2013 Dec;46(8):810-814 [FREE Full text] [doi: 10.1002/eat.22164] [Medline: 23881639]
46. Revelle W. psych: Procedures for Psychological, Psychometric, and Personality Research. version 2.2.5. The Comprehensive R Archive Network. Evanston, IL, USA: Northwestern University; 2022 Oct 14. URL: https://cran.r-project.org/web/packages/psych/psych.pdf [accessed 2022-11-15]
47. Dana J, Dawes RM. The superiority of simple alternatives to regression for social science predictions. J Educ Behav Stat 2016 Nov 23;29(3):317-331. [doi: 10.3102/10769986029003317]
48. Lichtenberg JM, Şimşek Ö. Simple regression models. In: Proceedings of the NIPS 2016 Workshop on Imperfect Decision Makers. 2016 Presented at: PMLR '16; December 9, 2016; Barcelona, Spain p. 13-25.
49. Arend AK, Kaiser T, Pannicke B, Reichenberger J, Naab S, Voderholzer U, et al. Code and Data for 'Toward Individualized Prediction of Binge-Eating Episodes Based on Ecological Momentary Assessment Data: Item Development and Pilot Study in Patients with Bulimia Nervosa and Binge-Eating Disorder'. Open Science Framework. 2022. URL: https://osf.io/p35y4/ [accessed 2022-12-16]
50. Dakanalis A, Carrà G, Calogero R, Fida R, Clerici M, Zanetti MA, et al. The developmental effects of media-ideal internalization and self-objectification processes on adolescents' negative body-feelings, dietary restraint, and binge eating. Eur Child Adolesc Psychiatry 2015 Aug;24(8):997-1010. [doi: 10.1007/s00787-014-0649-1] [Medline: 25416025]
51. Guertin TL, Conger AJ. Mood and forbidden foods' influence on perceptions of binge eating. Addict Behav 1999;24(2):175-193. [doi: 10.1016/s0306-4603(98)00049-5] [Medline: 10336100]

52. Heatherton TF, Polivy J, Herman CP. Dietary restraint: some current findings and speculations. Psychol Addict Behav 1990;4(2):100-106. [doi: 10.1037/h0080580]
53. Polivy J, Herman CP. Dieting and binging. A causal analysis. Am Psychol 1985 Feb;40(2):193-201. [doi: 10.1037//0003-066x.40.2.193] [Medline: 3857016]
54. Waterhouse J, Fukuda Y, Morita T. Daily rhythms of the sleep-wake cycle. J Physiol Anthropol 2012 Mar 13;31(1):5 [FREE Full text] [doi: 10.1186/1880-6805-31-5] [Medline: 22738268]
55. Bergmeir C, Hyndman RJ, Koo B. A note on the validity of cross-validation for evaluating autoregressive time series prediction. Comput Stat Data Anal 2018 Apr;120:70-83. [doi: 10.1016/j.csda.2017.11.003]
56. Breiman L. Random forests. Mach Learn 2001 Oct;45(1):5-32. [doi: 10.1023/A:1010933404324]
57. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016 Aug Presented at: KDD '16; August 13-17, 2016; San Francisco, CA, USA p. 785-794. [doi: 10.1145/2939672.2939785]
58. van der Laan MJ, Polley EC, Hubbard AE. Super learner. Stat Appl Genet Mol Biol 2007;6:Article25. [doi: 10.2202/1544-6115.1309] [Medline: 17910531]
59. Beck ED, Jackson JJ. Personalized prediction of behaviors and experiences: an idiographic person-situation test. Psychol Sci 2022 Oct;33(10):1767-1782. [doi: 10.1177/09567976221093307] [Medline: 36219572]
60. Pannicke B, Blechert J, Reichenberger J, Kaiser T. Clustering individuals' temporal patterns of affective states, hunger, and food craving by latent class vector-autoregression. Int J Behav Nutr Phys Act 2022 May 21;19(1):57 [FREE Full text] [doi: 10.1186/s12966-022-01293-1] [Medline: 35597952]
61. Forman EM, Goldstein SP, Zhang F, Evans BC, Manasse SM, Butryn ML, et al. OnTrack: development and feasibility of a smartphone app designed to predict and prevent dietary lapses. Transl Behav Med 2019 Mar 01;9(2):236-245 [FREE Full text] [doi: 10.1093/tbm/iby016] [Medline: 29617911]
62. Crochiere RJ, Zhang FZ, Juarascio AS, Goldstein SP, Thomas JG, Forman EM. Comparing ecological momentary assessment to sensor-based approaches in predicting dietary lapse. Transl Behav Med 2021 Dec 14;11(12):2099-2109. [doi: 10.1093/tbm/ibab123] [Medline: 34529044]
63. Juarascio AS, Crochiere RJ, Tapera TM, Palermo M, Zhang F. Momentary changes in heart rate variability can detect risk for emotional eating episodes. Appetite 2020 Sep 01;152:104698. [doi: 10.1016/j.appet.2020.104698] [Medline: 32278643]
64. Crochiere RJ, Kerrigan SG, Lampe EW, Manasse SM, Crosby RD, Butryn ML, et al. Is physical activity a risk or protective factor for subsequent dietary lapses among behavioral weight loss participants? Health Psychol 2020 Mar;39(3):240-244 [FREE Full text] [doi: 10.1037/hea0000839] [Medline: 31916827]
65. Walton A, Nahum-Shani I, Crosby L, Klasnja P, Murphy S. Optimizing digital integrated care via micro-randomized trials. Clin Pharmacol Ther 2018 Jul;104(1):53-58 [FREE Full text] [doi: 10.1002/cpt.1079] [Medline: 29604043]
66. Clyne C, Blampied NM. Training in emotion regulation as a treatment for binge eating: a preliminary study. Behav Change 2004 Dec 1;21(4):269-281. [doi: 10.1375/bech.21.4.269.66105]
67. Schag K, Rennhak SK, Leehr EJ, Skoda EM, Becker S, Bethge W, et al. IMPULS: impulsivity-focused group intervention to reduce binge eating episodes in patients with binge eating disorder - a randomised controlled trial. Psychother Psychosom 2019;88(3):141-153. [doi: 10.1159/000499696] [Medline: 31108488]

Abbreviations

AUC: area under the curve
BED: binge-eating disorder
BISCUIT: Best Item Scales that are Cross-validated, Unit-weighted, Informative and Transparent
BN: bulimia nervosa
CONSORT: Consolidated Standards of Reporting Trials
DSM-5: Diagnostic Statistical Manual-5
ED: eating disorder
EMA: ecologic momentary assessment
JITAI: just-in-time adaptive intervention

Edited by C Lovis; submitted 29.07.22; peer-reviewed by A Ruf, K Uludag; comments to author 06.09.22; revised version received 08.12.22; accepted 12.12.22; published 23.02.23.

Please cite as:
Arend AK, Kaiser T, Pannicke B, Reichenberger J, Naab S, Voderholzer U, Blechert J
Toward Individualized Prediction of Binge-Eating Episodes Based on Ecological Momentary Assessment Data: Item Development and Pilot Study in Patients With Bulimia Nervosa and Binge-Eating Disorder
JMIR Med Inform 2023;11:e41513
URL: https://medinform.jmir.org/2023/1/e41513
doi:10.2196/41513
PMID:36821359

Original Paper

Deployment of Real-time Natural Language Processing and Deep Learning Clinical Decision Support in the Electronic Health Record: Pipeline Implementation for an Opioid Misuse Screener in Hospitalized Adults

Majid Afshar[1], MSc, MD; Sabrina Adelaine[1], PhD; Felice Resnik[1], PhD; Marlon P Mundt[1], PhD; John Long[1], MS; Margaret Leaf[1], MS; Theodore Ampian[1], MS; Graham J Wills[1], PhD; Benjamin Schnapp[1], MS; Michael Chao[1], MS; Randy Brown[1], MD, PhD; Cara Joyce[2], PhD; Brihat Sharma[1], MS; Dmitriy Dligach[2], PhD; Elizabeth S Burnside[1], MPH, MD, MS; Jane Mahoney[1], MD; Matthew M Churpek[1], MD, PhD; Brian W Patterson[1], MPH, MD; Frank Liao[1], PhD

[1]University of Wisconsin - Madison, Madison, WI, United States

[2]Loyola University Chicago, Chicago, IL, United States

Corresponding Author:
Majid Afshar, MSc, MD
University of Wisconsin - Madison
1685 Highland Avenue
5158 Medical Foundation Centennial Building
Madison, WI, 53705
United States
Phone: 1 3125459462
Email: majid.afshar@wisc.edu

Abstract

Background: The clinical narrative in electronic health records (EHRs) carries valuable information for predictive analytics; however, its free-text form is difficult to mine and analyze for clinical decision support (CDS). Large-scale clinical natural language processing (NLP) pipelines have focused on data warehouse applications for retrospective research efforts. There remains a paucity of evidence for implementing NLP pipelines at the bedside for health care delivery.

Objective: We aimed to detail a hospital-wide, operational pipeline to implement a real-time NLP-driven CDS tool and describe a protocol for an implementation framework with a user-centered design of the CDS tool.

Methods: The pipeline integrated a previously trained open-source convolutional neural network model for screening opioid misuse that leveraged EHR notes mapped to standardized medical vocabularies in the Unified Medical Language System. A sample of 100 adult encounters were reviewed by a physician informaticist for silent testing of the deep learning algorithm before deployment. An end user interview survey was developed to examine the user acceptability of a best practice alert (BPA) to provide the screening results with recommendations. The planned implementation also included a human-centered design with user feedback on the BPA, an implementation framework with cost-effectiveness, and a noninferiority patient outcome analysis plan.

Results: The pipeline was a reproducible workflow with a shared pseudocode for a cloud service to ingest, process, and store clinical notes as Health Level 7 messages from a major EHR vendor in an elastic cloud computing environment. Feature engineering of the notes used an open-source NLP engine, and the features were fed into the deep learning algorithm, with the results returned as a BPA in the EHR. On-site silent testing of the deep learning algorithm demonstrated a sensitivity of 93% (95% CI 66%-99%) and specificity of 92% (95% CI 84%-96%), similar to published validation studies. Before deployment, approvals were received across hospital committees for inpatient operations. Five interviews were conducted; they informed the development of an educational flyer and further modified the BPA to exclude certain patients and allow the refusal of recommendations. The longest delay in pipeline development was because of cybersecurity approvals, especially because of the exchange of protected health information between the Microsoft (Microsoft Corp) and Epic (Epic Systems Corp) cloud vendors. In silent testing, the resultant pipeline provided a BPA to the bedside within minutes of a provider entering a note in the EHR.

Conclusions: The components of the real-time NLP pipeline were detailed with open-source tools and pseudocode for other health systems to benchmark. The deployment of medical artificial intelligence systems in routine clinical care presents an important yet unfulfilled opportunity, and our protocol aimed to close the gap in the implementation of artificial intelligence–driven CDS.

Trial Registration: ClinicalTrials.gov NCT05745480; https://www.clinicaltrials.gov/ct2/show/NCT05745480

(JMIR Med Inform 2023;11:e44977) doi:10.2196/44977

KEYWORDS

clinical decision support; natural language processing; medical informatics; opioid related disorder; opioid use; electronic health record; clinical note; cloud service; artificial intelligence; AI

Introduction

Background

As of 2017, >95% of the hospitals in the United States adopted an electronic health record (EHR), and >80% are collecting electronic clinical notes [1]. Clinical decision support (CDS) and intelligent data-driven alerts are part of federal incentive programs for Meaningful Use [2,3]. With the increasing capacity of EHR data and financial incentives to improve quality care, hospitals are increasingly well equipped to leverage computational resources to improve case identification and care throughput [4].

The unstructured narrative of EHRs provides a rich source of information on patients' conditions that may serve as CDS tools. Detailed medical information is routinely recorded in providers' intake notes. However, this information is neither organized nor prioritized during routine care for augmented intelligence at the bedside. Moreover, clinical notes' free-text format hinders efforts to perform analytics and leverage the large domain of data. The computational methods of natural language processing (NLP) can derive meaning from clinical notes, from which machine learning algorithms can screen for conditions such as opioid misuse.

In 2020, overdose deaths from opioid misuse soared to an all-time high, with a record 93,000 deaths nationwide during the pandemic year [5]. Substance misuse ranks second among the principal diagnoses for unplanned 7-day hospital readmission rates [6]. Screening for patients at risk for opioid use disorders is not part of the admission routine at many hospitals, and many hospitalized patients in need are never offered opioid treatment. The high prevalence rate of substance use disorders in hospitalized adults exceeds the rates in the general population or outpatient setting and reveals the magnitude of this lost opportunity [7]. We previously trained a convolutional neural network (CNN) that outperformed a rule-based approach and other machine learning methods for screening opioid misuse in hospitalized patients. The CNN substance misuse classifier had >80% sensitivity and specificity and demonstrated that clinical notes captured during hospitalization may be used to screen for opioid misuse [8].

There remains a paucity of evidence on the implementation of clinical NLP models in an interoperable and standardized CDS system for health operations and patient care [9]. The interactions among an artificial intelligence (AI) system, its users, its implementation, and the environment influence the AI intervention's overall potential effectiveness. Few health systems have been able to accommodate the complexities of an NLP deep learning model integrated into an existing operational ecosystem and EHR [10]. Much of the literature on NLP-driven CDS has described retrospective studies [11,12] outside the clinical workflow or simulated clinical environments [13,14]. Others have used NLP for information extraction efforts aimed at quality improvement without direct integration into the clinical workflow and operations [15,16]. Few provide a real-time NLP system but do not share an implementation framework or pipeline details to ensure fidelity and reproducibility [17]. Although the field of AI-driven CDS is growing, sharing knowledge in development and operations for health care delivery is lacking in best practices for processes and technologies in application planning, development, delivery, and operations.

This Study

This protocol describes a cloud service designed to ingest, process, and store clinical notes as standardized and interoperable messages from a major EHR vendor in an elastic cloud computing environment. We subsequently demonstrate the use of multiple open-source tools, including an open-source NLP engine for processing EHR notes and feeding them into a deep learning algorithm for screening for opioid misuse. Our resultant NLP and deep learning pipeline can process clinical notes and provide decision support to the bedside within minutes of a provider entering a note into the EHR.

To our knowledge, this is the first protocol for a bedside implementation of an NLP-driven CDS tool. We expect that our protocol will serve as a guide for other health systems to leverage open-source tools across interoperable data standards and ontologies. We provide an implementation framework and a cost-effectiveness analysis of a tool developed for the automated screening of hospitalized adults for opioid misuse. We aimed to describe a hospital-wide protocol and computing architecture for implementing a real-time NLP-driven CDS tool.

Methods

Hospital Setting and Study Period

The NLP CDS tool was implemented at the University of Wisconsin (UW) Hospital across the surgical and medical hospital inpatient wards. The EHR system used at the UW Health is Epic (Epic Systems Corp). The tool was designed for hospitalized adults (aged ≥18 years) and was assessed using a

pre-post quasi-experimental study design over 30 months (24 months of usual care and 6 months for the implementation of automated screening). The study was a quality improvement initiative by the health system to provide an automated hospital-wide screening system for opioid misuse and was registered on ClinicalTrials.gov (NCT05745480).

Preintervention Period: Usual Care With Ad Hoc Addiction Consultations

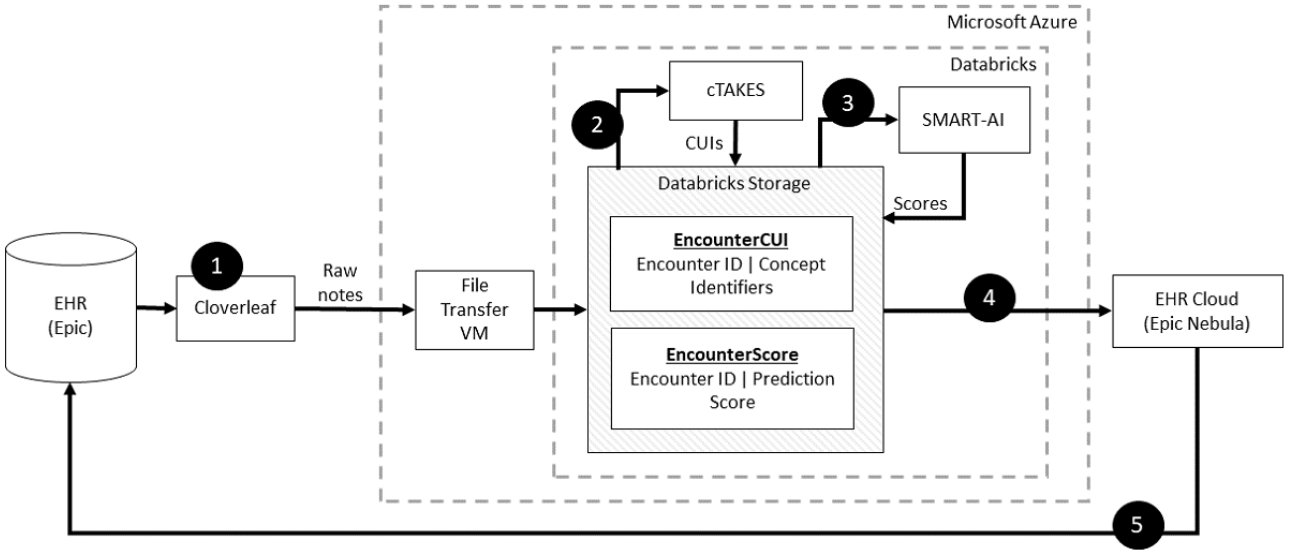
The UW Hospital launched an Addiction Medicine inpatient consult service in 1991 to address the high prevalence of substance use disorders among hospitalized adults. A screening, brief intervention, and referral to treatment program [18] was instituted for alcohol misuse. Screening, intervention flow sheets, and consult order sets were built into EHR-driven workflows for inpatient nurses and social workers for alcohol screening using the Alcohol Use Disorders Identification Test–Concise [19], a best practice alert (BPA) for patients at risk of alcohol use disorder, and order sets for withdrawal treatment. For other drugs, a single screening item queries "marijuana or other recreational drug use," but no formal screening process was in place specifically targeting opioid misuse. For patients at risk of an opioid use disorder, the practice was ad hoc consultations at the discretion of the primary provider.

Postintervention Period: Computing Architecture and Real-time Implementation

Overview

The technical architecture that enabled the real-time NLP CDS tool incorporated industry-leading and emerging technological capabilities. Figure 1 details the overall NLP CDS infrastructure that exported the notes from the EHR, organized them, and fed them into an NLP pipeline; input the processed text features into the opioid screener deep learning model; and delivered the resultant scores back to the bedside EHR as a BPA. The final architecture was a real-time NLP CDS tool, and the 6 components of the architecture are further detailed in the subsequent sections.

Figure 1. Step-by-step implementation of clinical natural language processing (NLP) pipeline. Step 1 ran a scheduled program to ingest notes from the EHR for each patient organized the notes, and relayed them via an HL7 data feed (Cloverleaf) into the cloud computing environment and data lake (Microsoft Azure and Databricks) onto a VM (Step 2). The NLP engine (cTAKES) processed the text stored on the VM and mapped them to medical concepts from the National Library of Medicine's metathesaurus (CUIs). The machine learning model received the CUIs as inputs and stored the results in DataBricks. At regular intervals, a custom Python script in Databricks performed the text extraction and linguistic feature engineering via cTAKES and stored CUIs with the appended data of patient identifiers. The CUIs served as the input to the machine learning model (SMART-AI) at the encounter level. The output of prediction probabilities and classification was stored in a Databricks table (Step 3). An API call from the EHR cloud is made to determine whether the cutpoint threshold from the machine learning model was met to trigger a BPA. In Step 4, the EHR cloud made an HTTP call to Databricks to request the score. The score was returned to the EHR cloud and subsequently delivered as a BPA when the provider opened the patient's chart in our on-premise instance of the EHR (Step 5). API: application programming interface; BPA: best practice alert; cTAKES: Clinical Text Analysis and Knowledge Extraction System; CUI: concept unique identifier; EHR electronic health record; HL7: Health Level 7; SMART-AI: Substance Misuse Algorithm for Referral to Treatment Using Artificial Intelligence; VM: virtual machine.



Component 1: Transferring Clinical Notes From the EHR to Cloud Computing

Health Level 7 (HL7) refers to the standards for transferring health care data between data sources. Cloverleaf (Infor Cloverleaf Integration Suite) was the UW's vendor solution that served as an application programming interface (API) gateway for accessing the clinical narratives in the EHR using HL7 version 2. To initialize the data feed, a UW Health interface analyst created a new entry in the Cloverleaf vendor software detailing the desired record information, which included clinical note text and identifiers. The analyst then "activated" the data feed, which began a continuous Transmission Control Protocol message–generating process. The Transmission Control Protocol messages were communicated using the HL7 application protocol to the Azure Virtual Machine (version 2022; Microsoft Corp) host at a port designated by the data science engineering team. This port was reserved by a NET program (the "TCP listener"), which wrote the message to the cloud file system and replied to the Cloverleaf server with acknowledgment messages,

conforming to the HL7 version 2 specification. Ultimately, the API extracted clinical notes from Epic and transferred them to a Microsoft Azure cloud computing environment that was under a business associate agreement with the UW. On-premise relays with the File Transfer Protocol were used to transfer the clinical notes to a specified location in the Azure cloud environment.

Component 2: Cloud Analytic Computing Platform

In the Microsoft Azure framework [8], the UW Health invoked the Databricks (Databricks Inc) analytic resources and services for scalable computing, data storage, and querying. The open-source tools from the NLP engine and our trained, publicly available machine learning model were hosted in Databricks. The machine learning model life cycle management (MLFlow) tool in Databricks supported the data flow for the deep learning model. MLFlow created and scored models when clinical notes were received and subsequently reported the results upon request. The final infrastructure was a scalable and failure-resistant environment for analytic computations.

Component 3: NLP Pipeline

The Clinical Text Analysis and Knowledge Extraction System (cTAKES; Apache Software Foundation) was built on multiple open-source Apache projects and incorporated technologies with the Unstructured Information Management Architecture framework and the Apache OpenNLP NLP toolkit [20]. This configuration contained several engines for sentence detection, tokenization, part-of-speech tagging, concept detection, and normalization to extract information from the clinical narrative in the EHR. We did not use the negation module because it was not used in the current use case; however, this can be turned on for other use cases. cTAKES is one of the most ubiquitous NLP engines used in the clinical domain [21]. cTAKES provided named entities from the free text that were mapped from the National Library of Medicine's Unified Medical Language System (UMLS), which is a repository of groups of words with relevant clinical contexts (eg, drugs, diseases, symptoms, anatomical sites, and procedures). Each named entity was mapped to a concept unique identifier (CUI) using the UMLS Systemized Nomenclature of Medicine–Clinical Terms and medical prescription normalized ontologies. For instance, "heroin misuse" from the text was assigned C0600241 as its CUI, which was different from the CUI assigned to "history of heroin misuse," C3266350. For generalizability, we used the default cTAKES pipeline [22].

As clinical notes were entered into the EHR for an individual patient, Cloverleaf relayed the notes via HL7 from the Epic EHR and used the Azure File Transfer Protocol server running on a virtual machine to place them at a known location within the Azure cloud environment. In 15-minute intervals, Databricks triggered a custom Python script to extract the text and fed it into the cTAKES pipeline to map and extract the CUIs. The CUIs were stored in the Azure Data Lake with appended data, including patient ID, encounter ID, and note time stamp, and were ready to be fed into any machine learning model. The code executed for the pipeline consisted of several services that operated independently and communicated through data stores. These services were "always on," but each had a trigger condition that initiated the code execution. The pseudocode for these services is provided in Multimedia Appendix 1.

Component 4: Text Feed From the NLP Pipeline Into the Deep Learning Model

We previously published a substance misuse screening algorithm using CUIs fed into a CNN called the Substance Misuse Algorithm for Referral to Treatment Using Artificial Intelligence (SMART-AI) [8]. SMART-AI was trained on the first 24 hours of all clinical notes entered into the EHR, starting from the patient's arrival time. This approach provided sufficient time not only for robust training but also for the addiction consult service to intervene before hospital discharge. For ease of implementation, the model was not trained on any specific note type and followed a time stamp approach for all notes filed within 24 hours of arrival at the hospital. SMART-AI is a supervised model with target labels that were derived from the manual screening data of over 50,000 patients who self-reported on the validated Drug Abuse Screening Test [23] and answered follow-up questions about opioid use. SMART-AI is publicly available to run the trained model [24], and more details about the model architecture and development can be found in the original development and validation publication [8]. The model's development and validation followed guideline recommendations [25].

Temporal validation of the classifier (trained on data between 2017 and 2019 and tested on data from 2020) at an outside hospital demonstrated an area under the precision-recall curve of 0.87 (95% CI 0.84-0.91) for screening for opioid misuse. Similar results were derived in an external validation at a second independent health system [8]. Multiple cutoff points were examined for the optimal threshold selection for the BPA, including the point on the area under the receiver operating curve that minimized the difference between sensitivity and specificity. During validation on the full cohort of hospitalized patients, the optimal cutoff point for screening for opioid misuse was 0.05. At that cutoff, the sensitivity was 0.87 (95% CI 0.84-0.90), specificity was 0.99 (95% CI 0·99-0·99), negative predictive value was 0.99 (95% CI 0.99-0.99), and positive predictive value was 0.76 (95% CI 0.72-0.88). The number needed to evaluate was 1.4, which translates to 26 alerts per 1000 hospitalized patients [8]. This was deemed an acceptable workload for consultation requests in live production for the UW Addiction Medicine clinicians. Additional silent testing was performed at the UW Health to examine sensitivity and specificity with 95% CI in our practice setting.

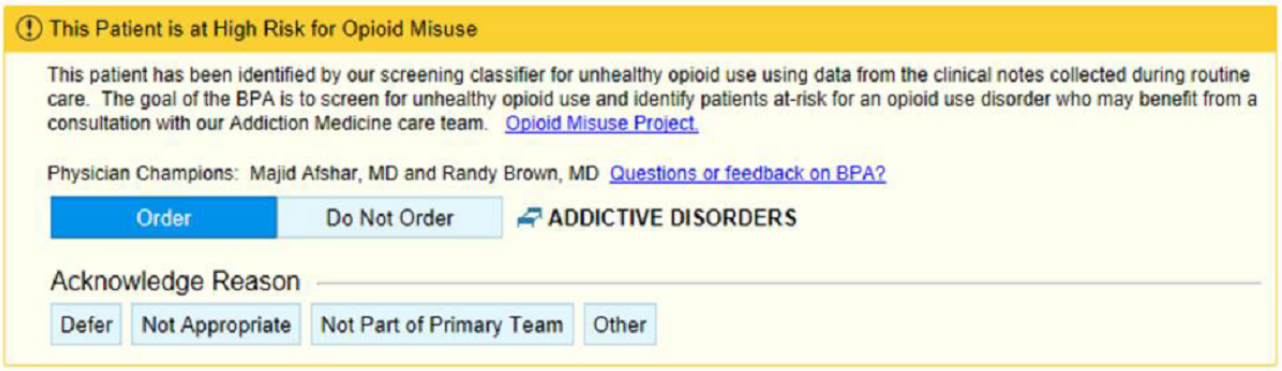
All notes from the first 24 hours of arrival at the UW Hospital were combined into a single document per patient encounter and converted into sequences of vector representations (eg, embeddings). The CUI embeddings defined the input layer to the SMART-AI model at the encounter level. The model provided prediction probabilities for opioid misuse and stored them in a Databricks table with the predefined cutoff point for screen positives.

Component 5: Real-time Delivery of the Prediction Results

The Nebula Cloud Platform was Epic's Software as a service platform for integrating new technology and specifically supported clinical prediction modeling. Nebula capabilities included the deployment of machine learning models, including a library of Epic-curated models for health care and custom algorithms. Our solution leveraged the latter to facilitate triggers from Epic to call out to the Databricks environment and provided the predictions for BPAs.

In the case of SMART-AI, we designed a BPA (Figure 2) to trigger once a clinician opened a patient chart in the EHR. Epic called its Nebula component to determine whether a BPA should be generated. Nebula made an HTTP call to Databricks to request the score. The RESTful HTTP API provided the SMART-AI model score that was serviced using MLFlow. The parameters included UMLS dictionaries, model results, patient identifiers, and other attributes necessary for individual-level predictions. The score was returned to Nebula, which was used to trigger a BPA if SMART-AI met the cutoff score for opioid misuse. For screen positives, the alert recommended the clinician to consult with the UW's Addiction Medicine consult service. The following were internal targets to meet the real-time needs of the end user at the bedside: (1) a throughput of 1000 notes per minute (<60 ms each); (2) three-nines (99.9%) availability—equivalent of <9 hours of downtime annually; and (3) an established error rate threshold.

Figure 2. In an iterative design with feedback from end users, a final BPA was implemented for bedside care. The BPA triggers upon opening a chart for a patient that meets the cutpoint predicted probability for opioid misuse from the NLP and deep learning model (SMART-AI). BPA: best practice alert; NLP: natural language processing; SMART-AI: Substance Misuse Algorithm for Referral to Treatment Using Artificial Intelligence.



Component 6: Cybersecurity

Two principles of security were applied: (1) defense in depth and (2) zero trust. The zero-trust architecture was outlined in the National Institute of Standards and Technology Special Publication 800 to 207 [26]. To secure access between Azure Databricks MLFlow and Epic's Nebula, we used an authentication token and IP range restriction (Databricks admin utility). The authentication token was issued via Databricks standard authentication. As a security best practice, we used the Databricks service principal and its Databricks access token to provide automated tool and system access to Databricks resources.

Implementation Framework

The Consolidated Framework for Implementation Research informed the development of the preimplementation assessments and will be used during the rapid Plan-Do-Study-Act (PDSA) cycles after deployment [27]. Key stakeholder interviews were planned to better understand the context and identify the barriers to and facilitators of the implementation of the BPA tool. Selected implementation strategies from the Expert Recommendations for Implementing Change were chosen to overcome barriers [28]. For pilot implementation, a regular cadence of meetings was planned with the implementation team to process, reflect, and evaluate the barriers to the implementation and use of the BPA. Process evaluation would incorporate interviews with providers and addiction specialists to understand what barriers still existed to using and acting on the BPA. During pilot implementation, we will collect and summarize clinical performance data during PDSA cycles to guide clinicians and administrators in monitoring, evaluating, and modifying provider behavior. Using the Consolidated Framework for Implementation Research–Expert Recommendations for Implementing Change matching tool [29], we will tailor relevant implementation strategies to enhance provider uptake and use of the tool. In addition, during the pilot phase, we will interview providers on the hospital units beyond the pilot units to identify and explore their determinants for the use of the BPA. After a pilot implementation period of 3 months, we will optimize provider training, enhance educational materials, and institute quality monitoring preparatory to a hospital-wide rollout.

Patient Outcome Analysis and Power Calculation

The SMART-AI study intervention sample consisted of all hospitalized patients who were screened positive for opioid misuse through the NLP CDS tool. The primary effectiveness measure was the percentage of hospitalized patients in the NLP CDS intervention sample who screened positive for opioid misuse and received an intervention by the inpatient addiction consult service. A control sample was derived by retrospectively applying the NLP CDS tool to all inpatient EHR records from the 2 years before this study's initiation in March 2023. Hospitalized patients who were screened positive retrospectively through the NLP CDS tool will form the usual care control group.

The primary outcome was the percentage of inpatients who were screened positive (or would have screened positive)

through the NLP CDS tool and who received an addiction consult with any of the following interventions: (1) receipt of opioid use intervention or motivational interviewing (MI), (2) receipt of medication-assisted treatment (MAT), or (3) referral to substance use disorder treatment. The primary outcome will be reported as a percentage in the preintervention and postintervention periods and will be measured through substance use screening and treatment service engagement for hospitalized patients screened for opioid misuse. The secondary outcomes included the 30-day unplanned hospital readmission rate. The criteria for unplanned hospital readmissions were adopted from the Centers for Medicare and Medicaid Services [30].

Hypothesis testing for intervention effects will be conducted using independent tests of the difference in the proportion of patients receiving MI, MAT, or referral to substance use disorder treatment. The null hypothesis was that the proportion of patients who screened positive and received any of the aforementioned interventions was lower (inferior) in the postintervention period than in the preintervention period, that is, H_0: $p_1 - p_2 \geq M$, where M denotes the noninferiority (eg, equivalence) margin, p_1 denotes the preperiod proportion, and p_2 denotes the postperiod proportion. The alternative 1-tailed test for noninferiority, that is, H_1: $p_1 - p_2 < M$, will be tested using the *Z* statistic. The noninferiority design was adopted to demonstrate that comprehensive screening may be as effective as manual screening but less costly via automated solutions. Our use case was an example of an AI system intended to improve efficiency and throughput within a reasonable timeframe for hospital operations. In these cases, statistically superior performance on outcomes may not be expected or required for prospective implementation, and interventions may be desirable if they are both substantially equivalent (noninferior) on clinical outcomes and cost-effective, given the high cost of building IT infrastructure, hiring vendors, and obtaining licensing and software support.

In hospital-wide screening, we expected a prevalence of 3% of adult inpatients with opioid misuse based on prior findings of hospital-wide analyses. A total sample size of 12,500 patients, with 10,000 in the preintervention 2-year period and 2500 in the postintervention 6-month period, had 85% power to detect a difference of +0.75% in the postintervention period (3.75%) compared with the preintervention period (3%), with a noninferiority difference of −0.5% using a 1-sided *Z* test with a significance level of 0.025.

Cost-effectiveness Analysis

Overview

Cost-effectiveness analysis will estimate the incremental costs of the SMART-AI intervention for the 6 months after the implementation compared with the 6 months before the implementation (ie, the added costs of the SMART-AI tool in reference to usual care) relative to the incremental effectiveness for the primary and secondary outcomes. The health economic evaluation would determine incremental intervention costs by examining the following: (1) the opportunity start-up costs of implementing the SMART-AI tool, (2) the incremental medical costs resulting from usual care for hospitalized patients with opioid misuse versus SMART-AI automated screening–supported care costs, and (3) the ongoing costs of administering and maintaining the SMART-AI tool.

The start-up costs of establishing SMART-AI substance use screening care would include the costs associated with developing and implementing the NLP CDS tool: (1) the cost of supporting the NLP and machine learning components and building the BPA in the EHR and (2) the cost of training the health professionals on tool use. The incremental costs between usual care and SMART-AI automated screening care were determined by calculating medical care costs before and after the implementation of SMART-AI. Medical costs associated with the hospitalization stay and all subsequent medical costs for the 30 days following hospital admission for the pre– and post–SMART-AI intervention periods were derived from hospital billing records and presented from the single-payer (a health system) perspective.

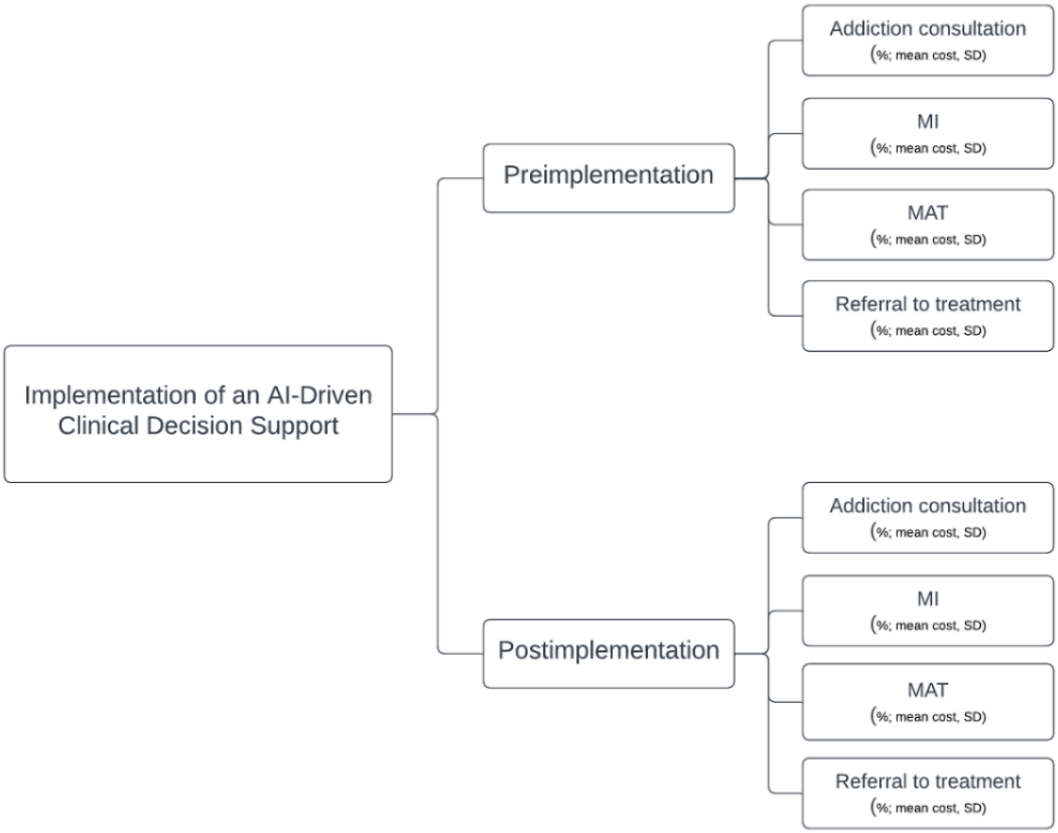
The following 3-pronged approach will be applied to identify the administration and maintenance costs associated with SMART-AI screening workflow changes introduced by the NLP CDS tool: (1) conducting in-depth interviews with hospital administrators, (2) performing activity-based observations of health care personnel who use SMART-AI, and (3) querying the clinician messaging system in the EHR. Average hospital compensation rates were used for valuing health care personnel time costs. Research-related costs were excluded.

Analytical Approach to Cost-effectiveness Analysis

The cost-effectiveness analysis was reported in terms of the incremental cost-effectiveness ratio (ICER) per additional patient who received substance use treatment. For this study, the ICER was calculated as the difference between preimplementation and postimplementation intervention costs divided by the difference between preimplementation and postimplementation intervention effectiveness as measured by the rates of patient engagement with substance use treatment services (ie, primary outcome) and 30-day hospital readmission (ie, secondary outcome).

The usual care control group and SMART-AI intervention group were characterized by the pathway probabilities of receiving substance use treatment and meeting the primary outcome. The pathway probabilities of patients' engagement with inpatient substance use consult, brief intervention or MI, MAT, and referral to substance use treatment for both study groups would result in 8 treatment combinations, which are displayed in Figure 3.

Figure 3. Cost-effectiveness decision tree. AI: artificial intelligence; MAT: medication-assisted treatment; MI: motivational interviewing.



The differential costs pre– and post–SMART-AI intervention were determined as the difference in the weighted sum of the individual pathway costs, using the pathway probabilities as weights for the intervention and control groups. Effectiveness was determined as the difference in the rates of hospitalized patients engaging with substance use disorder treatment before and after the implementation of SMART-AI for the intervention and control groups. The ICER was calculated as follows:

Sensitivity analyses will introduce uncertainty in substance use treatment receipt rates and costs for the intervention and control groups. The Monte Carlo–based simulation estimation used the rates of substance use treatment service uptake observed in the intervention and control groups as a reference to simulate a cohort of postimplementation hospitalized patients and a cohort of usual care hospitalized patients. The ICER per additional individual who received an inpatient substance use consult, brief intervention, MI, MAT, or referral to substance use treatment was calculated by drawing a random sample with replacement from the observed distributions for health care costs (μ_{COSTi}) and substance use treatment services (μ_{TRTi}) for the intervention and control groups. This process was repeated (n=1000) to produce bootstrap estimates of the 95% CI for the ICER per additional individual who received an inpatient substance use consult, brief intervention, MI, MAT, or referral to substance use treatment. These probabilistic sensitivity analyses estimated the elasticity of the differential cost per patient relative to the differential substance use treatment service rates for the intervention and control groups.

Ethics Approval

This clinical study was reviewed by the UW Institutional Review Board (ID 2022-0384). The study was part of a larger quality improvement initiative at the UW Health and met the exemption status for human participant research according to the UW Institutional Review Board. The study was secondary research with the collection of existing EHR data that met category 4 exemption. The study met the requirements for a waiver of consent, and all study results were anonymized or deidentified. No compensation was provided in the human participant research.

Results

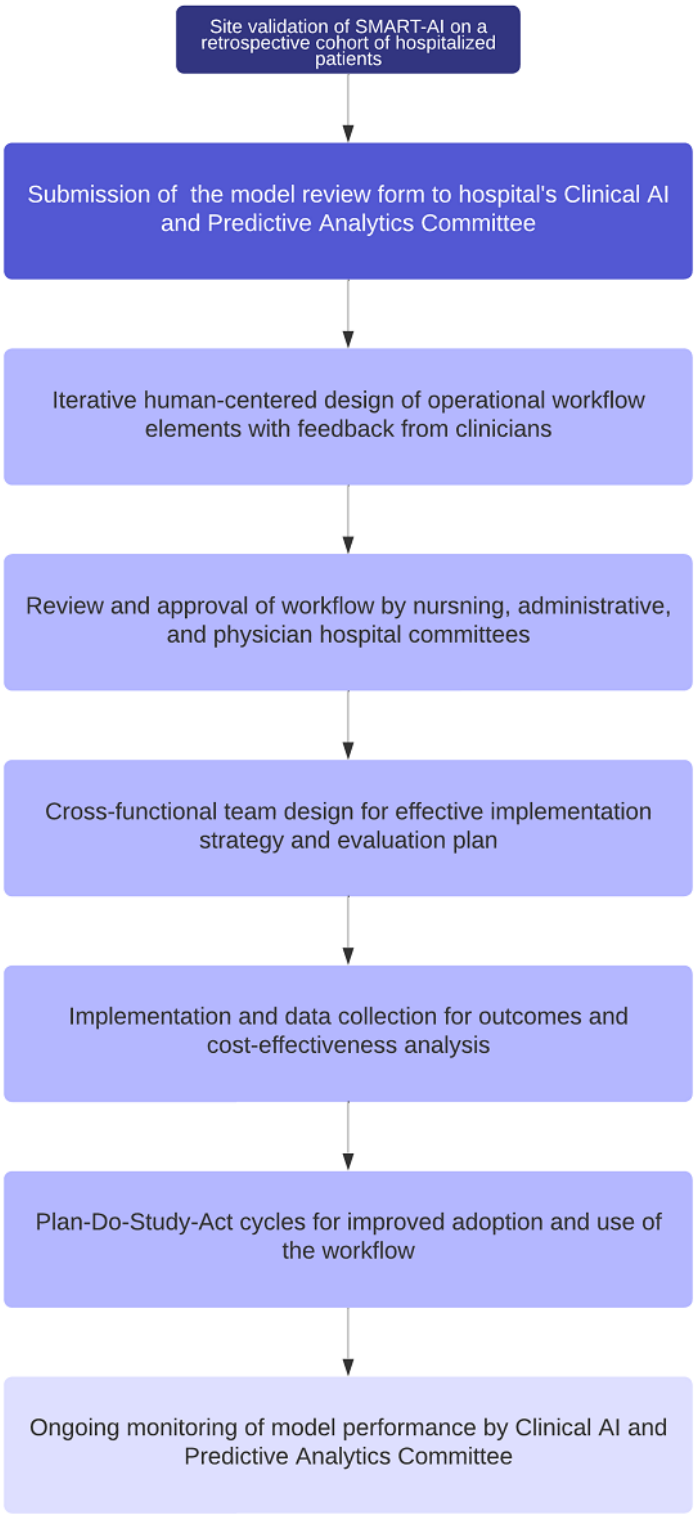
Preimplementation Testing and Approvals

Early-stage investigations were performed to assess the AI system's predictive performance in a retrospective setting and evaluate the human factors surrounding the BPA before initiating the quasi-experimental clinical study. During the silent testing of SMART-AI at the UW Health, a random sample of 100 adult patient encounters (with an oversampling of patients with the International Classification of Diseases codes for substance use) in 2021 were extracted and reviewed by an inpatient physician and a clinical informatics expert. SMART-AI performed similarly to previously published reports for screening for opioid misuse, with a sensitivity of 93% (95% CI 66%-99%) and specificity of 92% (95% CI 84%-96%).

Before the deployment of SMART-AI, approvals were received across hospital committees for inpatient operations, EHR super users, CDS, and nursing documentation. The proposal protocol

was also reviewed by the Center for Clinical Knowledge Management to confirm that there were no competing interests or roles with existing protocols for screening for substance use conditions in the health system. In addition, SMART-AI was reviewed by the UW's Clinical AI and Predictive Analytics Committee. A model review form providing details on the clinical problem, model value proposition, model description, proposed workflow integration, internal validation, and monitoring strategy (including fairness and equity) was reviewed and approved by a multidisciplinary committee of clinicians, informaticians, bioethicists, executive leadership, and data scientists. The planned workflow from introduction to implementation is shown in Figure 4.

Figure 4. Flow diagram for the process to bedside implementation and evaluation. AI: artificial intelligence; SMART-AI: Substance Misuse Algorithm for Referral to Treatment Using Artificial Intelligence.



Implementation Framework

An end-user interview guide and survey were developed to examine the user acceptability of the BPA. Open-ended questions were asked about the barriers to and facilitators of the use of the BPA. A total of 5 interviews were conducted (with 3 nurse practitioners, 1 family medicine resident, and 1 surgical attendant), and the responses led the production team to create an educational flyer, modify the BPA with more details and options for consultation refusal, and modify when and where the BPA would trigger. Figure 2 shows the final production version of the BPA for deployment. Dissemination efforts included Grand Round presentations to the Addiction Medicine Division, Department of Family Medicine, and notification via the hospital's weekly electronic newsletter.

The longest delay in operational workflow and architecture was for receiving cybersecurity approvals, especially for the exchange of protected health information between the Microsoft and Epic cloud vendors. An additional 6 months of delay occurred for achieving acceptable security monitors and checks. The go-live of SMART-AI in the EHR was scheduled for January 2023.

Discussion

Principal Findings

We offer one of the first protocols that detailed the components of a real-time NLP-driven CDS system for health care delivery at the bedside. We further detailed an implementation framework with human-centered design principles and a planned iterative process to evaluate the cost-effectiveness and health outcomes of screening for opioid misuse. We shared the components and pseudocode with open-source technologies involved in the implementation of an end-to-end NLP pipeline that processed the notes entered by the provider and returned a BPA within minutes for patients at risk of an opioid use disorder. Interviews and user-centered design as well as educational efforts for improving adherence led to changes in the BPA. Finally, we shared an experimental design with a rapid PDSA cycle and cost-effectiveness setup with a noninferiority design to evaluate the screening system for continued implementation or deimplementation.

The digital era in medicine continues to grow exponentially in terms of both the quantity of unstructured data collected in the EHR and the number of prediction models developed for detection and diagnostic, prognostic, and therapeutic guidance. In parallel, the clinical NLP field has grown in its capabilities with the advent of transformer architectures and more affordable and efficient cognitive computing of big data [31]. However, a major bottleneck remains in the successful implementation of NLP and deep learning models in clinical practice. Much of the progress in NLP has focused on information retrieval and extraction [32]; however, the application of these methods at scale with a combination of software developers and operations remains challenging at health care institutions. The role of NLP in BPAs has been limited to date, and prior BPAs have used existing technologies embedded into the EHR [33]. Similar to prior motivations for BPAs delivered to bedside clinicians [34], our intention was to support and enhance decision-making at the beside with a recommendation for an Addiction Medicine consult in patients who may otherwise not receive it or have it delayed, similar to another NLP-driven BPA [17]. However, given the lack of capacity of many EHR vendors to incorporate custom NLP models, we offer an interoperable pipeline to integrate external AI tools with existing EHRs.

Applied clinical NLP has predominantly remained a rule-based approach, but statistical machine learning models are now the leading method in the research literature [21]. Few vendors who provide NLP services rely entirely on machine learning, and a gap remains in effectively applying NLP models to EHRs that go beyond disease detection, which is limited to explicit keyword mentions [35]. Several barriers exist with neural language models, including the need to remove protected health information so that the trained models may be shared and the computational requirements to run complex deep learning models in a production environment [9]. We offer solutions for both barriers using a feature engineering approach to map free text to coded vocabulary and describe a large computing infrastructure with a connection between a data science cloud platform and the EHR to support direct data feeds into any machine learning model. The NLP CDS pipeline accomplishes efficiency in data standardization and scalability [36] for successful implementation and is extensible to other NLP engines. The benefit of augmented intelligence remains unknown and its identification using our health care outcomes and cost-effectiveness analysis is the next step in a clinical study.

Our implementation framework is largely guided by a team of implementation scientists supported by the UW's Clinical and Translational Science Award. We leveraged our Clinical and Translational Science Award's Dissemination and Implementation Launchpad to help bridge the gap between evidence-based research and practice [37]. The Dissemination and Implementation Launchpad serves to accelerate the pace of disseminating research findings and increase the adoption and implementation of effective interventions, leading to sustainable practice and policy changes. It uses strategies from implementation science, design thinking, and human-centered engineering for the better integration of AI technologies into health systems. As part of the preimplementation phase, we assessed contextual factors that may impact implementation by engaging both adopters, who are the decision-makers, and end users, who are the main implementers, of the tool [38]. We conducted qualitative interviews with end users to evaluate the need for the tool and BPA design. We involved adopters early in the process to inform the intervention or implementation process through consultations during the design, feasibility testing, and implementation phases. An iterative process ensued to address the constraints and contextual factors that affect the adoption and implementation of the tool in our health system.

During the preimplementation phase, the project team clarified roles with the project management, with the readiness of the clinical workflow approved through hospital committee meetings and individual interviews with end users. Our health system is an early adopter of AI governance with a review process similar to that of other health systems [39]. The Clinical AI and Predictive Analytics Committee follows the Minimum

Information About Clinical AI Modeling checklist [40]. The offline validation of our model incorporated principles from multiple reporting guidelines on prediction models, bias, fairness, and validation [41]. Clinical evaluation after the go-live of SMART-AI will follow the reporting guideline for the early-stage clinical evaluation of decision support systems driven by AI (Developmental and Exploratory Clinical Investigations of Decision Support Systems Driven by AI) [42].

The build of an enterprise-wide AI infrastructure for data-driven CDS is an important feature of a data-driven learning health system. At the UW, learning health system activities dating back to 2013 established an evidence-based framework with a series of organizational-level quality improvement interventions [43]. In 2020, the UW Health reaffirmed its strategic plan for embedding discovery and innovation as well as diversity, equity, and inclusion in clinical care. Successful implementation included coaching staff and administrative leaders for working in PDSA with lean management to get the problem, analysis, corrective actions, and action plan down on a single sheet of large (A3) paper, also known as “A3” thinking [44]. A rapid PDSA cycle is important in the advent of AI-driven interventions that require rigorous evaluation for implementation or deimplementation. Furthermore, the pipeline developed for the opioid screener use case is applicable to other CDS tools that use machine learning and NLP. We designed our architecture to ingest different modalities of data and provide a computing environment that is flexible to different data modalities and machine learning algorithms.

Several limitations exist in the deployment and sustainability of our NLP-driven CDS tool. First, calibration drift is a real concern with changes in medical practice, evidence, and demographic shifts over time that may affect model performance [45]. During implementation, reviews by the Clinical AI and Predictive Analytics Committee will include quarterly evaluations of the sustained effectiveness of the tool, an audit of the fairness of the tool across parity groups, and examination for alert fatigue. Others have shown benefits in recalibration approaches and domain adaptation with additional training data to update the models over time [46]. Furthermore, the start-up costs of the pipeline may be cost-prohibitive for small health systems. Our proposed cost-effectiveness analysis will provide a perspective on both the start-up costs of implementing the NLP tool and the ongoing incremental costs. The start-up costs are more of a burden to a small health system than the incremental costs, but we expect that our results will be informative in terms of both these costs.

Conclusions

The deployment of medical AI systems in routine clinical care presents an important yet unfulfilled opportunity [47], and our protocol aims to close the gap in the implementation of AI-driven CDS. Our protocol implementation for an enterprise-wide production environment of an AI opioid misuse screener provides a model for other health systems to use to bring NLP models into practice for CDS. We highlight opportunities to leverage the expertise of our applied data science team to use the open-source tools for feature engineering and model development inside a larger infrastructure with vendor support for hardware and software dependencies. Given the sensitive nature of health care data, the biggest challenges are ensuring high standards for cybersecurity and meeting the privacy requirements for protecting patient data.

Acknowledgments

The authors acknowledge support from the University of Wisconsin Institute for Clinical and Translational Research, which, in turn, is supported by the Clinical and Translational Science Award program through the National Institutes of Health National Center for Advancing Translational Sciences grant (2UL1TR002373). The research was also supported by the National Institute on Drug Abuse of the National Institutes of Health (R01DA051464; CJ, DD, MA, and RB), National Library of Medicine Temporal Histories of Your Medical Event (THYME) project (R01LM010090; DD), and National Institute of Diabetes and Digestive and Kidney Diseases R01DK126933). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the other funding sources listed in this section.

Data Availability

The raw electronic health record data are not available because of regulatory and legal restrictions imposed by the University of Wisconsin-Madison Institutional Review Board. The original data were derived from the institution’s electronic health record and contain patients’ protected health information. Data are available from the University of Wisconsin Health Systems for researchers who meet the criteria for access to confidential data and have a data use agreement with the health system. Only the final trained model that is fully deidentified with a vocabulary of mapped concept unique identifiers is open source and available [38]. Our deidentification approach has been previously described [39].

Authors' Contributions

MA, FL, BWP, SA, FR, CJ, MPM, DD, and MMC led the conception and design of the study and supervised the study. MA, FL, MPM, TA, GJW, and MC had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. CJ, B Sharma, and DD could not access the original data directly because of limitations in the data use agreement but take responsibility for the accuracy of the data analysis and did have access to all the data presented in the manuscript. MA, SA, MPM, ML, TA, GJW, B Schnapp, MC, CJ, B Sharma, ESB, FR, JM, DD, BWP, MMC, and FL performed

the analysis or interpretation of data. Administrative, technical, and material support were provided by MA, BWP, SA, JL, TA, GJW, B Schnapp, and FL. All authors reviewed the manuscript and provided edits and revisions. All authors take responsibility for the integrity of the work as a whole, from inception to the finished article, and all authors approved the final version submitted. MA was responsible for the decision to submit the manuscript.

Conflicts of Interest

Research conducted by RB is supported by grants to the University of Wisconsin by the Heffter Research Institute, Usona Institute, Revive Therapeutics, and the Etheridge Foundation. MMC is a named inventor on a patent for a risk stratification algorithm for hospitalized patients (US patent #11,410,777).

Multimedia Appendix 1
Pseudocode for custom Python scripts.
[DOCX File , 16 KB - medinform_v11i1e44977_app1.docx]

References

1. Adler-Milstein J, Holmgren AJ, Kralovec P, Worzala C, Searcy T, Patel V. Electronic health record adoption in US hospitals: the emergence of a digital "advanced use" divide. J Am Med Inform Assoc 2017 Nov 01;24(6):1142-1148 [FREE Full text] [doi: 10.1093/jamia/ocx080] [Medline: 29016973]
2. Meaningful use. Center for Disease Control and Prevention. 2022 Oct 3. URL: https://www.cdc.gov/vaccines/programs/iis/meaningful-use/index.html [accessed 2022-10-14]
3. Lite S, Gordon WJ, Stern AD. Association of the meaningful use electronic health record incentive program with health information technology venture capital funding. JAMA Netw Open 2020 Mar 02;3(3):e201402 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.1402] [Medline: 32207830]
4. Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ 2020 Mar 25;368:m689 [FREE Full text] [doi: 10.1136/bmj.m689] [Medline: 32213531]
5. Brown KG, Chen CY, Dong D, Lake KJ, Butelman ER. Has the United States reached a plateau in overdoses caused by synthetic opioids after the onset of the COVID-19 pandemic? Examination of centers for disease control and prevention data to November 2021. Front Psychiatry 2022 Jul 07;13:947603 [FREE Full text] [doi: 10.3389/fpsyt.2022.947603] [Medline: 35873233]
6. Fingar KR, Barrett ML, Jiang HJ. A comparison of all-cause 7-day and 30-day readmissions, 2014. In: Healthcare Cost and Utilization Project (HCUP) Statistical Briefs. Rockville, MD, USA: Agency for Healthcare Research and Quality; Oct 2017.
7. Owens PL, Fingar KR, McDermott KW, Muhuri PK, Heslin KC. Inpatient stays involving mental and substance use disorders, 2016. In: Healthcare Cost and Utilization Project (HCUP) Statistical Briefs. Rockville, MD, USA: Agency for Healthcare Research and Quality; May 08, 2019.
8. Afshar M, Sharma B, Dligach D, Oguss M, Brown R, Chhabra N, et al. Development and multimodal validation of a substance misuse algorithm for referral to treatment using artificial intelligence (SMART-AI): a retrospective deep learning study. Lancet Digit Health 2022 Jun;4(6):e426-e435 [FREE Full text] [doi: 10.1016/S2589-7500(22)00041-3] [Medline: 35623797]
9. Lederman A, Lederman R, Verspoor K. Tasks as needs: reframing the paradigm of clinical natural language processing research for real-world decision support. J Am Med Inform Assoc 2022 Sep 12;29(10):1810-1817 [FREE Full text] [doi: 10.1093/jamia/ocac121] [Medline: 35848784]
10. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med 2018 May 8;1:18 [FREE Full text] [doi: 10.1038/s41746-018-0029-1] [Medline: 31304302]
11. Van Vleck TT, Chan L, Coca SG, Craven CK, Do R, Ellis SB, et al. Augmented intelligence with natural language processing applied to electronic health records for identifying patients with non-alcoholic fatty liver disease at risk for disease progression. Int J Med Inform 2019 Sep;129:334-341 [FREE Full text] [doi: 10.1016/j.ijmedinf.2019.06.028] [Medline: 31445275]
12. Hu D, Li S, Zhang H, Wu N, Lu X. Using natural language processing and machine learning to preoperatively predict lymph node metastasis for non-small cell lung cancer with electronic medical records: development and validation study. JMIR Med Inform 2022 Apr 25;10(4):e35475 [FREE Full text] [doi: 10.2196/35475] [Medline: 35468085]
13. Park EH, Watson HI, Mehendale FV, O'Neil AQ, Clinical Evaluators. Evaluating the impact on clinical task efficiency of a natural language processing algorithm for searching medical documents: prospective crossover study. JMIR Med Inform 2022 Oct 26;10(10):e39616 [FREE Full text] [doi: 10.2196/39616] [Medline: 36287591]
14. Blackley SV, MacPhaul E, Martin B, Song W, Suzuki J, Zhou L. Using natural language processing and machine learning to identify hospitalized patients with opioid use disorder. AMIA Annu Symp Proc 2021 Jan 25;2020:233-242 [FREE Full text] [Medline: 33936395]

15. Rybinski M, Dai X, Singh S, Karimi S, Nguyen A. Extracting family history information from electronic health records: natural language processing analysis. JMIR Med Inform 2021 Apr 30;9(4):e24020 [FREE Full text] [doi: 10.2196/24020] [Medline: 33664015]
16. Tamang S, Humbert-Droz M, Gianfrancesco M, Izadi Z, Schmajuk G, Yazdany J. Practical considerations for developing clinical natural language processing systems for population health management and measurement. JMIR Med Inform 2023 Jan 03;11:e37805 [FREE Full text] [doi: 10.2196/37805] [Medline: 36595345]
17. Smith JC, Spann A, McCoy AB, Johnson JA, Arnold DH, Williams DJ, et al. Natural language processing and machine learning to enable clinical decision support for treatment of pediatric pneumonia. AMIA Annu Symp Proc 2021 Jan 25;2020:1130-1139 [FREE Full text] [Medline: 33936489]
18. Vaca FE, Winn D. The basics of alcohol screening, brief intervention and referral to treatment in the emergency department. West J Emerg Med 2007 Aug;8(3):88-92 [FREE Full text] [Medline: 19561690]
19. Bradley KA, DeBenedetti AF, Volk RJ, Williams EC, Frank D, Kivlahan DR. AUDIT-C as a brief screen for alcohol misuse in primary care. Alcohol Clin Exp Res 2007 Jul;31(7):1208-1217. [doi: 10.1111/j.1530-0277.2007.00403.x] [Medline: 17451397]
20. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc 2010;17(5):507-513 [FREE Full text] [doi: 10.1136/jamia.2009.001560] [Medline: 20819853]
21. Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, et al. Clinical information extraction applications: a literature review. J Biomed Inform 2018 Jan;77:34-49 [FREE Full text] [doi: 10.1016/j.jbi.2017.11.011] [Medline: 29162496]
22. Bleeker T. Welcome to the Apache cTAKES documentation. cTAKES. 2023 Feb 7. URL: https://cwiki.apache.org/confluence/display/CTAKES/ [accessed 2022-07-13]
23. Yudko E, Lozhkina O, Fouts A. A comprehensive review of the psychometric properties of the Drug Abuse Screening Test. J Subst Abuse Treat 2007 Mar;32(2):189-198. [doi: 10.1016/j.jsat.2006.08.002] [Medline: 17306727]
24. Substance misuse algorithm for referral to treatment using artificial intelligence (SMART-AI). GitHub. 2021 Nov 23. URL: https://github.com/Rush-SubstanceUse-AILab/SMART-AI [accessed 2022-07-24]
25. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med 2015 Jan 06;162(1):55-63 [FREE Full text] [doi: 10.7326/M14-0697] [Medline: 25560714]
26. Argaw ST, Troncoso-Pastoriza JR, Lacey D, Florin MV, Calcavecchia F, Anderson D, et al. Cybersecurity of hospitals: discussing the challenges and working towards mitigating the risks. BMC Med Inform Decis Mak 2020 Jul 03;20(1):146 [FREE Full text] [doi: 10.1186/s12911-020-01161-7] [Medline: 32620167]
27. Damschroder LJ, Aron DC, Keith RE, Kirsh SR, Alexander JA, Lowery JC. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. Implement Sci 2009 Aug 07;4:50 [FREE Full text] [doi: 10.1186/1748-5908-4-50] [Medline: 19664226]
28. Powell BJ, Waltz TJ, Chinman MJ, Damschroder LJ, Smith JL, Matthieu MM, et al. A refined compilation of implementation strategies: results from the Expert Recommendations for Implementing Change (ERIC) project. Implement Sci 2015 Feb 12;10:21 [FREE Full text] [doi: 10.1186/s13012-015-0209-1] [Medline: 25889199]
29. Waltz TJ, Powell BJ, Fernández ME, Abadie B, Damschroder LJ. Choosing implementation strategies to address contextual barriers: diversity in recommendations and future directions. Implement Sci 2019 Apr 29;14(1):42 [FREE Full text] [doi: 10.1186/s13012-019-0892-4] [Medline: 31036028]
30. Zuckerman RB, Sheingold SH, Orav EJ, Ruhter J, Epstein AM. Readmissions, observation, and the hospital readmissions reduction program. N Engl J Med 2016 Apr 21;374(16):1543-1551. [doi: 10.1056/NEJMsa1513024] [Medline: 26910198]
31. Laparra E, Mascio A, Velupillai S, Miller T. A review of recent work in transfer learning and domain adaptation for natural language processing of electronic health records. Yearb Med Inform 2021 Aug;30(1):239-244 [FREE Full text] [doi: 10.1055/s-0041-1726522] [Medline: 34479396]
32. Gao Y, Dligach D, Christensen L, Tesch S, Laffin R, Xu D, et al. A scoping review of publicly available language tasks in clinical natural language processing. J Am Med Inform Assoc 2022 Sep 12;29(10):1797-1806. [doi: 10.1093/jamia/ocac127] [Medline: 35923088]
33. Chaparro JD, Beus JM, Dziorny AC, Hagedorn PA, Hernandez S, Kandaswamy S, et al. Clinical decision support stewardship: best practices and techniques to monitor and improve interruptive alerts. Appl Clin Inform 2022 May;13(3):560-568. [doi: 10.1055/s-0042-1748856] [Medline: 35613913]
34. Chen H, Butler E, Guo Y, George Jr T, Modave F, Gurka M, et al. Facilitation or hindrance: physicians' perception on best practice alerts (BPA) usage in an electronic health record system. Health Commun 2019 Aug;34(9):942-948. [doi: 10.1080/10410236.2018.1443263] [Medline: 29485296]
35. Chiticariu L, Li Y, Reiss FR. Rule-based information extraction is dead! long live rule-based information extraction systems!. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. 2013 Presented at: EMNLP '13; October 18-21, 2013; Seattle, WA, USA p. 827-832.
36. Asan O, Choudhury A. Research trends in artificial intelligence applications in human factors health care: mapping review. JMIR Hum Factors 2021 Jun 18;8(2):e28236 [FREE Full text] [doi: 10.2196/28236] [Medline: 34142968]

37. Mehta TG, Mahoney J, Leppin AL, Stevens KR, Yousefi-Nooraie R, Pollock BH, et al. Integrating dissemination and implementation sciences within Clinical and Translational Science Award programs to advance translational research: recommendations to national and local leaders. J Clin Transl Sci 2021 Jul 12;5(1):e151 [FREE Full text] [doi: 10.1017/cts.2021.815] [Medline: 34527291]
38. Rolland B, Resnik F, Hohl SD, Johnson LJ, Saha-Muldowney M, Mahoney J. Applying the lessons of implementation science to maximize feasibility and usability in team science intervention development. J Clin Transl Sci 2021 Jul 22;5(1):e197 [FREE Full text] [doi: 10.1017/cts.2021.826] [Medline: 34888066]
39. Liao F, Adelaine S, Afshar M, Patterson BW. Governance of clinical AI applications to facilitate safe and equitable deployment in a large health system: key elements and early successes. Front Digit Health 2022 Aug 24;4:931439 [FREE Full text] [doi: 10.3389/fdgth.2022.931439] [Medline: 36093386]
40. Norgeot B, Quer G, Beaulieu-Jones BK, Torkamani A, Dias R, Gianfrancesco M, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. Nat Med 2020 Sep;26(9):1320-1324 [FREE Full text] [doi: 10.1038/s41591-020-1041-y] [Medline: 32908275]
41. Collins GS, Dhiman P, Andaur Navarro CL, Ma J, Hooft L, Reitsma JB, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. BMJ Open 2021 Jul 09;11(7):e048008 [FREE Full text] [doi: 10.1136/bmjopen-2020-048008] [Medline: 34244270]
42. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, DECIDE-AI expert group. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med 2022 May;28(5):924-933. [doi: 10.1038/s41591-022-01772-9] [Medline: 35585198]
43. Kraft S, Caplan W, Trowbridge E, Davis S, Berkson S, Kamnetz S, et al. Building the learning health system: describing an organizational infrastructure to support continuous learning. Learn Health Syst 2017 Oct;1(4):e10034 [FREE Full text] [doi: 10.1002/lrh2.10034] [Medline: 31245569]
44. Lee TS, Kuo MH. Toyota A3 report: a tool for process improvement in healthcare. Stud Health Technol Inform 2009;143:235-240. [Medline: 19380942]
45. Siregar S, Nieboer D, Versteegh MI, Steyerberg EW, Takkenberg JJ. Methods for updating a risk prediction model for cardiac surgery: a statistical primer. Interact Cardiovasc Thorac Surg 2019 Mar 01;28(3):333-338. [doi: 10.1093/icvts/ivy338] [Medline: 30608590]
46. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 2010 Jan;21(1):128-138 [FREE Full text] [doi: 10.1097/EDE.0b013e3181c30fb2] [Medline: 20010215]
47. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med 2022 Jan;28(1):31-38. [doi: 10.1038/s41591-021-01614-0] [Medline: 35058619]

Abbreviations

AI: artificial intelligence
API: application programming interface
BPA: best practice alert
CDS: clinical decision support
CNN: convolutional neural network
cTAKES: Clinical Text Analysis and Knowledge Extraction System
CUI: concept unique identifier
EHR: electronic health record
HL7: Health Level 7
ICER: incremental cost-effectiveness ratio
MAT: medication-assisted treatment
MI: motivational interviewing
MLFlow: machine learning model life cycle management
NLP: natural language processing
PDSA: Plan-Do-Study-Act
SMART-AI: Substance Misuse Algorithm for Referral to Treatment Using Artificial Intelligence
UMLS: Unified Medical Language System
UW: University of Wisconsin

Edited by G Eysenbach, C Perrin; submitted 11.12.22; peer-reviewed by L Tong, D Chrimes, P Han; comments to author 11.01.23; revised version received 01.02.23; accepted 26.03.23; published 20.04.23.

Please cite as:
Afshar M, Adelaine S, Resnik F, Mundt MP, Long J, Leaf M, Ampian T, Wills GJ, Schnapp B, Chao M, Brown R, Joyce C, Sharma B, Dligach D, Burnside ES, Mahoney J, Churpek MM, Patterson BW, Liao F
Deployment of Real-time Natural Language Processing and Deep Learning Clinical Decision Support in the Electronic Health Record: Pipeline Implementation for an Opioid Misuse Screener in Hospitalized Adults
JMIR Med Inform 2023;11:e44977
URL: https://medinform.jmir.org/2023/1/e44977
doi:10.2196/44977
PMID:37079367

Original Paper

Standardized Comparison of Voice-Based Information and Documentation Systems to Established Systems in Intensive Care: Crossover Study

Arne Peine[1,2], MHBA, MD; Maike Gronholz[2]; Katharina Seidl-Rathkopf[2], PhD; Thomas Wolfram[2], MBA, MD; Ahmed Hallawa[1], MSc; Annika Reitz[2], MSc; Leo Anthony Celi[3,4], MD, PhD; Gernot Marx[1], Prof Dr; Lukas Martin[1,2], MD, PhD

[1]Department of Intensive Care Medicine and Intermediate Care, University Hospital RWTH Aachen, Aachen, Germany

[2]Clinomic Group GmbH, Aachen, Germany

[3]Laboratory of Computational Physiology, Harvard–MIT Division of Health Sciences Technology, Cambridge, MA, United States

[4]Beth Israel Deaconess Medical Center, Boston, MA, United States

Corresponding Author:
Arne Peine, MHBA, MD
Department of Intensive Care Medicine and Intermediate Care
University Hospital RWTH Aachen
Pauwelsstr. 30
Aachen, 52070
Germany
Phone: 49 241 800
Email: apeine@ukaachen.de

Abstract

Background: The medical teams in intensive care units (ICUs) spend increasing amounts of time at computer systems for data processing, input, and interpretation purposes. As each patient creates about 1000 data points per hour, the available information is abundant, making the interpretation difficult and time-consuming. This data flood leads to a decrease in time for evidence-based, patient-centered care. Information systems, such as patient data management systems (PDMSs), are increasingly used at ICUs. However, they often create new challenges arising from the increasing documentation burden.

Objective: New concepts, such as artificial intelligence (AI)–based assistant systems, are hence introduced to the workflow to cope with these challenges. However, there is a lack of standardized, published metrics in order to compare the various data input and management systems in the ICU setting. The objective of this study is to compare established documentation and retrieval processes with newer methods, such as PDMSs and voice information and documentation systems (VIDSs).

Methods: In this crossover study, we compare traditional, paper-based documentation systems with PDMSs and newer AI-based VIDSs in terms of performance (required time), accuracy, mental workload, and user experience in an intensive care setting. Performance is assessed on a set of 6 standardized, typical ICU tasks, ranging from documentation to medical interpretation.

Results: A total of 60 ICU-experienced medical professionals participated in the study. The VIDS showed a statistically significant advantage compared to the other 2 systems. The tasks were completed significantly faster with the VIDS than with the PDMS (1-tailed t_{59}=12.48; Cohen d=1.61; P<.001) or paper documentation (t_{59}=20.41; Cohen d=2.63; P<.001). Significantly fewer errors were made with VIDS than with the PDMS (t_{59}=3.45; Cohen d=0.45; P=.03) and paper-based documentation (t_{59}=11.2; Cohen d=1.45; P<.001). The analysis of the mental workload of VIDS and PDMS showed no statistically significant difference (P=.06). However, the analysis of subjective user perception showed a statistically significant perceived benefit of the VIDS compared to the PDMS (P<.001) and paper documentation (P<.001).

Conclusions: The results of this study show that the VIDS reduced error rate, documentation time, and mental workload regarding the set of 6 standardized typical ICU tasks. In conclusion, this indicates that AI-based systems such as the VIDS tested in this study have the potential to reduce this workload and improve evidence-based and safe patient care.

(JMIR Med Inform 2023;11:e44773) doi:10.2196/44773

KEYWORDS

artificial intelligence; documentation; ICU; intensive care medicine; speech-recognition; user perception; workload

Introduction

Overview

Intensive care medicine is the interdisciplinary treatment of patients with critical illnesses in specialized wards called intensive care units (ICUs) [1,2]. Patients admitted to ICUs have complex courses of disease and related treatments. However, ensuring adequate and safe measures in the ICU is often difficult due to a combination of short stays among patients, a high cognitive workload, and a limited number of rotating staff members [3]. The staff's time distribution is crucial for patient care and treatment. The direct contact between doctors and patients plays a crucial role in patient and family satisfaction as well as physicians' work satisfaction [4].

Patient contact, including treatment and communication, only makes up about one-fifth of the doctors' work time, while about one-third of their time is spent on documentation and data interpretation [5]. The documentation for patients with critical illnesses is complex and therefore labor-intensive [2]. This leads to an abundance of information with a high density of data arising from different bedside devices, consequently making their interpretation difficult and time-intensive [3]. Thus, documentation and evaluation tasks make up an increasing part of the physician's work time and become a major part of work on ICUs [6].

Physicians report that, due to rising requirements in documentation, they are under constant time pressure and complain about lacking time for patient-centered care [7]. This increasing burden is one of the crucial driving forces for burnout syndrome in physicians [8-10]. Similarly, a central part of the nurses' work on ICUs is the collection and documentation of massive, however critical, amounts of data and information about their patients [11]. This information overload and the redundancy in documentation can impair the ability to recognize the development of critical situations early on [3,12]. In response to this increasing workload, an improved documentation system is needed to enable the ICU staff to focus on their patients and decrease the time spent on documentation.

Workload in ICUs

Several aspects contribute to the workload in ICUs. First, medical knowledge in general is growing exponentially [13]. Medical guidelines and treatments increase in dynamicity and complexity as medical decisions and treatments at the patient's bed need to follow the current state of scientific evidence [14]. Additionally, the amount of generated health data doubles every 3 years [15]. Currently, an ICU patient produces more than 1000 data points per hour [15]. Thus, complex patient monitoring often leads to an information overload with unstructured and context-free data [3].

This makes it difficult to extract the most significant (and thus decision-relevant) aspects of a patient's history and the course of the disease [7,16]. The 2011 study by Ahmed et al [17] proves that the way in which the large amount of data generated in ICUs are presented has an impact on the viewer's ability to put it into the correct context. The more data are presented to the viewer, the higher the associated error rate. The phenomenon is even aggravated by the introduction of electronic health records and patient data management systems (PDMSs), especially when they present complete, unfiltered data sets [17]. Often, these systems therefore draw the immediate attention of the treating staff, resulting in less focus on the actual patient treatment [18].

These phenomena even have an imminent impact on the staff's occupational and mental health. It has been demonstrated that the more time physicians spend on less satisfying tasks (such as documentation), the higher the risk for burnout syndrome will become [19]. Shanafelt et al [19] showed in their 2009 analysis that the most important factor for burnout in physicians was to spend less than 20% of their time on the most meaningful activity (odds ratio 2.75, 95% CI 1.13-4.6; $P<.001$). The association between a physician's subjective work experience and the quality of patient care was already underlined in a 1985 study by Grol et al [20]. A similar finding for nursing staff was made in 2019 by Manomenidis et al [21]. They proved that, connected to burnout syndrome, the hand hygiene compliance in nursing staff was decreasing significantly, leading to reduced patient safety.

Documentation Systems in ICUs

Documentation facilitates interdisciplinary information flow and enhances continuity in patient care [22]. There are different forms of documentation systems available for this purpose in an ICU setting, such as traditionally used paper-based documentation, electronic health records with PDMSs, and new software developments that make use of speech recognition and artificial intelligence (AI).

Paper documentation has the advantage of being cost-efficient and simple to use, as no IT infrastructure has to be implemented [23]. However, disadvantages include the lack of on-site documentation, the lack of simultaneous access, and redundancy and ineffectiveness in documentation [24]. PDMS solutions have been developed to replace paper files, coordinate records from bedside equipment and laboratories, and thus reduce the ICU team's workload [25].

Many studies have investigated the pros and cons of PDMS implementation and show heterogeneous outcomes. Some studies underline that electronic health records lead to time savings, uniformity, and readability of the documentation. Ubiquitous and parallel availability of the patient's files reduces idle time and minimizes interruptions in documentation and data assessment [26,27]. Hence, information flow can be increased while the error rate can be decreased [3]. On the contrary, other studies show that PDMS generates a large amount of data, which, depending on the presentation, can make it difficult to identify relevant data and increase error rates, thus leading to a higher workload [7,17,28,29].

Currently, as many studies have proven, there is an imbalance between the time spent on completing documentation tasks and direct patient care [4,5,7,11,30]. Several studies show that the steady increase in time spent on documentation can be traced back to the introduction of electronic health records [7,25,27,31].

Speech recognition, a technique mainly driven by AI, can support the completion of documentation tasks. It is commonly used in consumer hardware devices and has been proven to increase productivity while reducing costs in the medical domain [28,32]. The use of computerized voice recognition in the medical domain is currently being investigated [33]. Several studies showed a reduction in documentation time with the use of speech recognition [29,33-36]. The reports produced offered greater word variety and more detailed and longer texts [34]. Additionally, an accelerated information flow and an improved subjective efficiency were proven [32,36].

Nevertheless, speech recognition is not well-established; thus, several studies examining the aforementioned technology showed a high error rate with significantly more critical clinical errors [28,35]. The main reasons for the errors in speech recognition are the use of nonnative speakers, difficulties in recognizing medical terms, and the ambient noise common in an ICU setting [33,34].

Speech recognition technologies have the potential to reduce the workload, especially regarding documentation tasks. However, it is necessary to develop a system that addresses the issues currently found in the use of speech recognition in order to establish the technology in clinical settings. This study investigates the use of an AI-based voice recognition technology for typical ICU tasks.

The development and introduction of a new documentation system must be based on the challenges in current technologies. A new system has to be intuitive and easy to understand, especially as the introduction period is often perceived as an addition burden since increased time has to be spent on the same tasks [27,36].

In addition, a new system should autonomously record and summarize patient data [11,37]. Thus, the vast amount of data produced by an ICU patient should be recorded, saved, and organized in the new system automatically, without the work of the medical team, in order to reduce workload. Flohr et al [3] confirm in their study that the automated collection of all patient data—the ability to view it in summarized form, identify trends, and have clear patient lists—can facilitate decisions and reduce workload and error rates [3].

While on the one hand, speech recognition can possibly reduce the documentational burden, on the other hand, AI can enable a well-structured, relevance-oriented patient presentation and clinical decision support [30,38].

Although of high importance for clinical outcomes, limited studies have been performed to compare new AI-based, voice-controlled documentation and information software with established documentation systems for ICUs (eg, paper documentation and classical PDMS computerized input) [32,34]. In this crossover study, we compare performance, mental workload, documentation accuracy, and user experience between methods. The objective of this study is to compare established documentation and retrieval processes with newer methods, such as PDMSs and voice information and documentation systems (VIDSs). This study compares performance, documentation accuracy, mental workload, and user experience associated with 6 tasks typical of the ICU as they are completed using 3 different approaches (paper-based, PDMS, and VIDS).

Methods

Study Design

Material

The study design includes 3 different ICU documentation tools.

An Established PDMS (IntelliSpace Critical Care and Anesthesia, Version J.01.00; Koninklijke Philips N.V.)

IntelliSpace Critical Care and Anesthesia is a clinical documentation and decision support system that includes a flowsheet, calculations engine, clinical advisories, device interfacing, orders, microbiology and pathology results, dietary and nursing orders, order management, infusion management, and numerous other functionalities [39]. We generated a fictitious patient using medical data mimicking a typical intensive care patient, including laboratory values, findings, demographic data, and other clinically relevant aspects. The validity of the data was confirmed by 2 independent trained intensivists.

Paper-Based, Conventional Documentation on Patient Curves (MEDLINQ Curve, Version 03.17)

In order to reflect the complexity of an intensive care patient, the participants received paper-based documentation (“patient curve”) of the same fictitious patient. This record included laboratory values, care reports, an anesthesia protocol, microbiology requests, a document concerning the patient’s belongings, and the patient curve. The curve consisted of 4 pages used to document the patient status for 1 day. The curve was completed for the patient until 6 AM, which was the time the participants were asked to complete the tasks. The biggest emphasis was placed on the fact that the paper-based sheet contained the same clinical information as was included in the other arms.

VIDS (Mona, Version P1.2; Clinomic GmbH)

In order to equalize potential differences between the study arms as much as possible, the AI-based software was installed on a portable system-on-chip computer (NVIDIA AGX Xavier) so that all study assessments could be performed in the same location. The system was based on adapted, proprietary software (“Mona”). For the purpose of the study, an adapted version of the Mona system was used, containing the following components: (1) voice handling capabilities: natural language understanding and processing; (2) data processing and preparation algorithms; (3) user interface components; and (4) voice synthesis components. The system was running on a Linux-based operation system and was further equipped with a directional microphone (Bose VideoMic NTG microphone) in order to create optimal voice recording circumstances (Figure 1) [40]. The VIDS interacted with an electronic health record

for each patient. The system was able to extract and display information from the patient data, enter patient care–related tasks, and navigate through charts. The user activated the system by saying, "Hey Mona." The system then played a short sound to reflect that voice recognition was activated. Throughout the interaction between the system and the user, the conversation was displayed for the user to read. The system would ask the user for any missing information to complete a task or answer a question. The patient information was shown in the form of tables and abstracts from the flow sheets. After a task was completed or a question was answered, the user had to reactivate the voice recognition with the words "Hey Mona."

Figure 1. Overview of the voice information and documentation system [40].

Procedure

A total of 60 medical professionals were included in the study. To achieve a sufficient level of significance, the size of the necessary sample was calculated as follows: sample size = $[z^2 \times p(1-p)]/e^2/1 + [z^2 \times p(1-p)]/e^2 \times N]$, where N = population size, z = *z* score, e = margin of error, and p = SD. The assumptions here were as follows: the total number of health care workers in Germany is 5.8 million, of whom 10% are working in ICUs. Confidence level of 0.85, resulting in a *z* score of 1.65 and a margin of error of 0.1. As a result, the required sample size was set at n>52. The definition of "ICU-experienced" included trained physicians, medical students after their fourth study year, and ICU nurses. The participants were recruited between February 5, 2021, and May 14, 2021. Participants were recruited in different domains of ICUs in order to represent the widest possible range of potential users of an information and documentation system. Participation was voluntary and could be terminated at any time without consequences.

First, each participant was informed about all parts of the study and then asked to sign an informed consent form. This was followed by the completion of a questionnaire about demographics, professional background, and technical affinity. The latter was measured using the validated "Fragebogen zur Erfassung der Technikaffinität als Umgang mit und Einstellung zu elektronischen Geräte" (TA-EG, "Questionnaire for the assessment of technology affinity as handling and attitude toward electronic devices") questionnaire [41] (Multimedia Appendix 1).

The participants then moved over to the study location. In order to reproduce the high noise level of ICUs during the COVID-19 pandemic, the tests were performed in a noisy simulation environment where other people were simultaneously working and moving around. The participants were asked to work on 6 different tasks typical in ICU workflow. The first 3 of the tasks were documentation tasks, followed by an assessment of a patient's status (either lactate trend or creatinine levels), and the last one was the generation of the ICU-relevant score "sequential organ failure assessment." Details on the given tasks are presented in Textbox 1. This list of tasks had to be completed with each of the examined systems mentioned above. Details about the questionnaires and methodology used can be found in Multimedia Appendix 2. The order in which the participants were presented with the different study arms was randomized before the start of the study using the randomize functionality of Excel (Microsoft Corp). The crossover design was chosen to ensure the comparability of all 3 interventions with respect to confounding variables. In addition, it allowed the risk of first- and second-type errors to be kept as low as possible when needing to minimize the number of necessary participants during the COVID-19 pandemic. While the participant executed the

tasks, the time needed was measured for each system. Each participant completed the tasks with the respective systems in all study arms. The order of steps in the study can be seen in Figure 2. The detailed solution pathway for each task completion can be seen in Multimedia Appendix 3.

Textbox 1. Tasks presented to the study participants.

Documentation

- "Document 300 mg amiodarone IV now"
- "Document 1.5 g piperacillin/tazobactam (Tazobac) intravenously now/at 10:00 AM"
- "Document the administration of a red blood cell concentrate/fresh frozen plasma at 1:00 AM for procedures numbered 1101002233 and indication active bleeding"
- "Document 20 mg furosemide now"

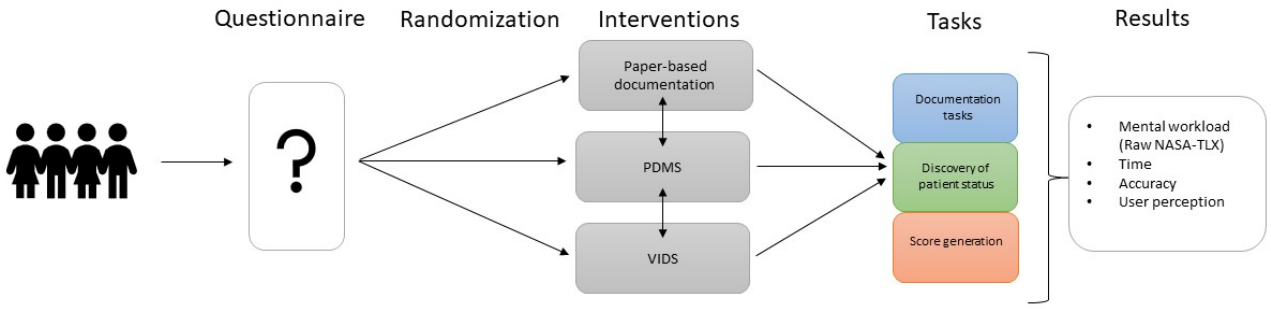
Discovery of patient status

- "What was the lactate trend in the last 12 hours?" Or "how did the creatinine levels develop within the last 4 days?"

Score generation

- "What is the patient's current SOFA score?"

Figure 2. Study procedure. NASA-TLX: National Aeronautics and Space Administration Task Load Index; PDMS: patient data management system; VIDS: voice information and documentation system.



After completing all tasks for 1 study branch, the participants filled out the German translation of the Raw National Aeronautics and Space Administration Task Load Index (NASA-TLX; Raw Task Load Index [RTLX]), a 6-item questionnaire that assesses subjective workload in 6 subscales without using a weighted ranking [42,43]. The shorter version of the NASA-TLX has great use in industry and research due to its simplicity while being equivalent to the task load index, and is thus recommended as a workload assessment tool [44]. For each participant, a mean value was calculated from the answers to the 6 subscales. A high RTLX score correlates with a high mental workload.

Lastly, the participants completed a final questionnaire to assess subjective user perception. This included the meCUE2.0, which is a questionnaire measuring user experience [45]. The meCUE2.0 was filled out for every system (Multimedia Appendix 2). The answers to the 6-point Likert scale were converted into numerical values so that the number 6 represents a maximum positive user experience. Additionally, the net promoter score (NPS) for each system was assessed. This score is a market research metric correlating with actual consumer behavior. It analyzes how likely a company or system is to be recommended by dividing the participants into detractors, neutrals, and promoters [46].

All questionnaires were filled out on a tablet using a publicly accessible, web-based survey system (LimeSurvey, version 3.23.3). Participants' responses were recorded for each task and subsequently compared against a clinically validated gold standard to assess accuracy. The gold standard was determined by 2 independent ICU specialists. A third independent ICU specialist was consulted in case of disagreement between the 2 physicians to determine the solution for the gold standard. Each error was counted as a negative point. Errors were defined as responses that deviated from the independently generated gold standard or tasks that were not completed.

There was no involvement of patients; all the data used were anonymous and fictitious.

Statistical Analysis

All statistical preprocessing and analysis were carried out using Excel and SPSS Statistics 27 (version 27.0.0.0; IBM Corp). For all statistical procedures, the α level was set at .05. The results were rounded to 2 decimal digits. According to the central limiting value theorem, an approximative normal distribution was assumed as the sample size was more than 30 [47]. The Levene test was used for the homoscedasticity requirement. A repeated, nonparametric, 1-way ANOVA was used in order to examine potential statistical differences between the different study arms.

Ethical Considerations

This study was approved by the ethics commission of the Medical Faculty of Rheinisch-Westfaelische Technische Hochschule Aachen (EK370/19).

Results

Overview

A total of 60 participants were included in the study; no dropouts occurred during the course of the study. Of the 60 participants, 26 identified as male and 34 as female. The average age was 32.87 (SD 12.46) years. The age ranged from 21 to 63 years. The groups were represented by 43% (26/60) physicians, 40% (24/60) students, and 17% (10/60) ICU nurses. Within their respective professional groups, 32% (19/60) reported 3-5 years of work experience, 27% (16/60) >10 years, and 18% (11/60) 5-10 years. Overall, 17% (10/60) had <1 year of work experience, and 7% (4/60) of participants reported experience of 1-3 years. In total, 6 participants reported having a native language other than German. Overall, 63% (38/60) of the 60 participants worked in a top-level hospital (>500 beds), 25% (15/60) in an upper-level hospital (300-500 beds), and 12% (7/60) in a primary care hospital (up to 300 beds). The technology affinity score was assessed with a mean of 3.70 (SD 0.47). The detailed characteristics of the study can be found in Table 1.

Table 1. Study population.

Variable	Value (n=60)
Age (years), mean (SD)	32.87 (12.46)
Gender, n (%)	
Female	34 (57)
Male	26 (43)
Diverse	0 (0)
Work setting, n (%)	
Top-level hospital (>500 beds)	38 (63)
Upper-level hospital (300-500 beds)	15 (25)
Primary care hospital (≤300 beds)	7 (12)
Professional group, n (%)	
Physician	26 (43)
ICU[a] Nurse	10 (17)
Medical student	24 (40)
Work experience (years), n (%)	
<1	10 (17)
1-3	4 (7)
3-5	19 (32)
5-10	11 (18)
>10	16 (27)

[a]ICU: intensive care unit.

Objective Parameters

The results of the analysis of the objective parameters—accuracy, time, and mental workload—showed that the participants performed best using the VIDS prototype. The exact results can be seen in Table 2. The differences between the systems were then analyzed using repeated measures ANOVA with a Greenhouse-Geisser correction.

Table 2. Descriptive statistics and objective parameters.

System	Accuracy, mean (SD)	Time (seconds), mean (SD)	RTLX[a], mean (SD)
VIDS[b]	0.28 (0.52)	195.45 (79.08)	3.49 (1.52)
PDMS[c]	0.75 (0.88)	491.03 (177.63)	4.17 (1.62)
Paper	2.37 (1.46)	763.93 (242.14)	6.31 (1.71)

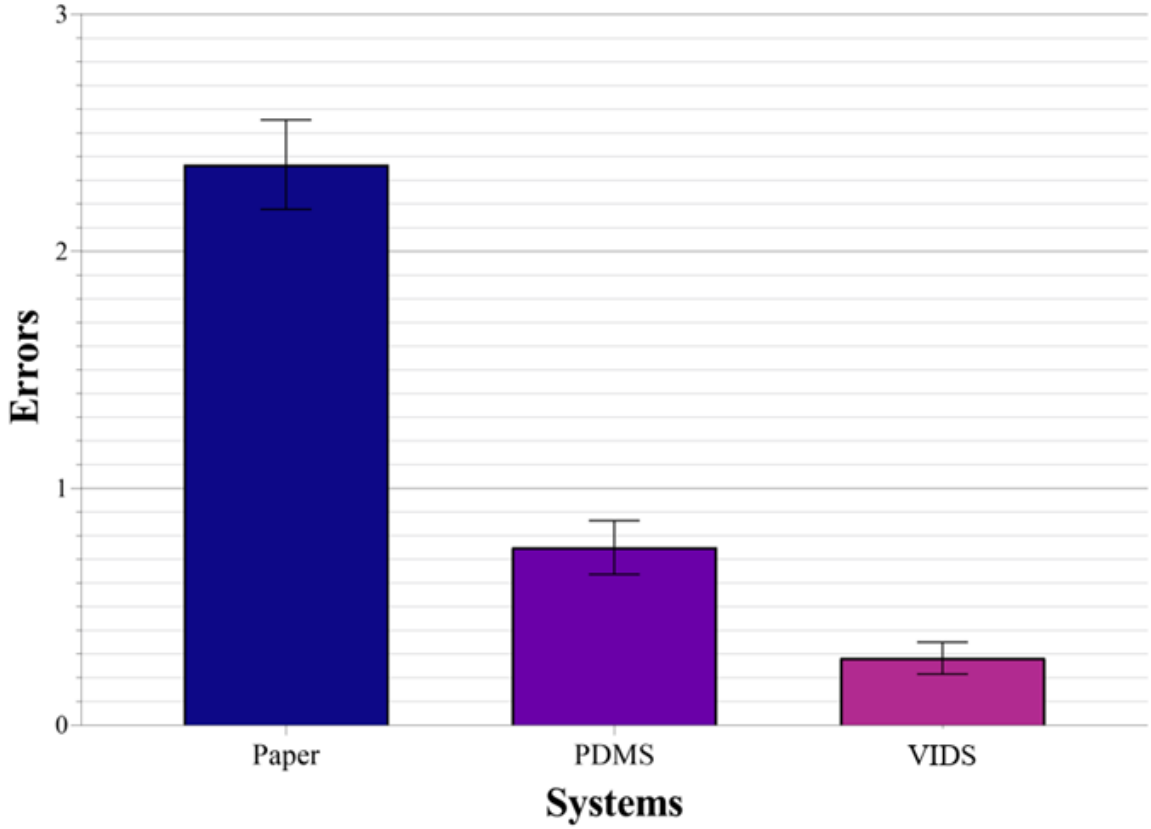
[a]RTLX: Raw Task Load Index.

[b]VIDS: voice information and documentation system.

[c]PDMS: patient data management system.

The analysis determined that there is a statistically significant difference in the errors made (n=60; $F_{1.73, 101.77}$=78.92; η_p^2=0.57; *P*<.001). The statistically significant difference is visualized in Figure 3. The effect size is Cohen *d*=1.15 and thus shows a large effect [48]. Post hoc analysis with a Bonferroni adjustment revealed that by using the VIDS, significantly fewer errors were made compared to PDMS (1-tailed t_{59}=3.45; Cohen *d*=0.45; *P*=.03) to paper-based documentation (t_{59}=11.2; Cohen *d*=1.45; *P*<.001) [49]. Using PDMS, significantly fewer errors were made compared to paper-based documentation (t_{59}=8.3; Cohen *d*=1.07; *P*<.001) [49].

Figure 3. Mean errors with standard error of the mean. PDMS: patient data management system; VIDS: voice information and documentation system.



Similar results were seen in the analysis of the time needed by the participants when completing the tasks, as can be seen in Figure 4. A statistically significant difference between the 3 systems was proven (n=60; $F_{1.65, 97.25}$=188.84; η_p^2=0.76; Cohen *d*=1.79; *P*<.001) [48]. The post hoc analysis with Bonferroni adjustments confirmed that the use of the VIDS was statistically significantly faster (VIDS to PDMS: t_{59}=12.48; Cohen *d*=1.61; *P*<.001; VIDS to paper: t_{59}=20.41; Cohen *d*=2.63; *P*<.001; PDMS to paper: t_{59}= 7.78; Cohen *d*=1; *P*<.001) [49].

The repeated measures ANOVA of the mental workload also showed a statistically significant difference between the use of the 3 systems (n=60; $F_{1.82, 107.11}$=56.91; η_p^2=0,49; Cohen *d*=0.98; *P*<.001), as shown in Figure 5 [48]. However, the post hoc analysis with Bonferroni adjustments only proved a statistically significant higher mental workload using paper-based documentation compared to PDMS (t_{59}=9.27; Cohen *d*=1.2; *P*<.001) and VIDS (t_{59}=9.16; Cohen *d*=1.18; *P*<.001) [49]. There is no statistically significant reduction in workload when using VIDS compared to PDMS (t_{59}=2.4; *P*=.06) [49].

Figure 4. Mean measured time in between the study arms with standard error of the mean. PDMS: patient data management system; VIDS: voice information and documentation system.

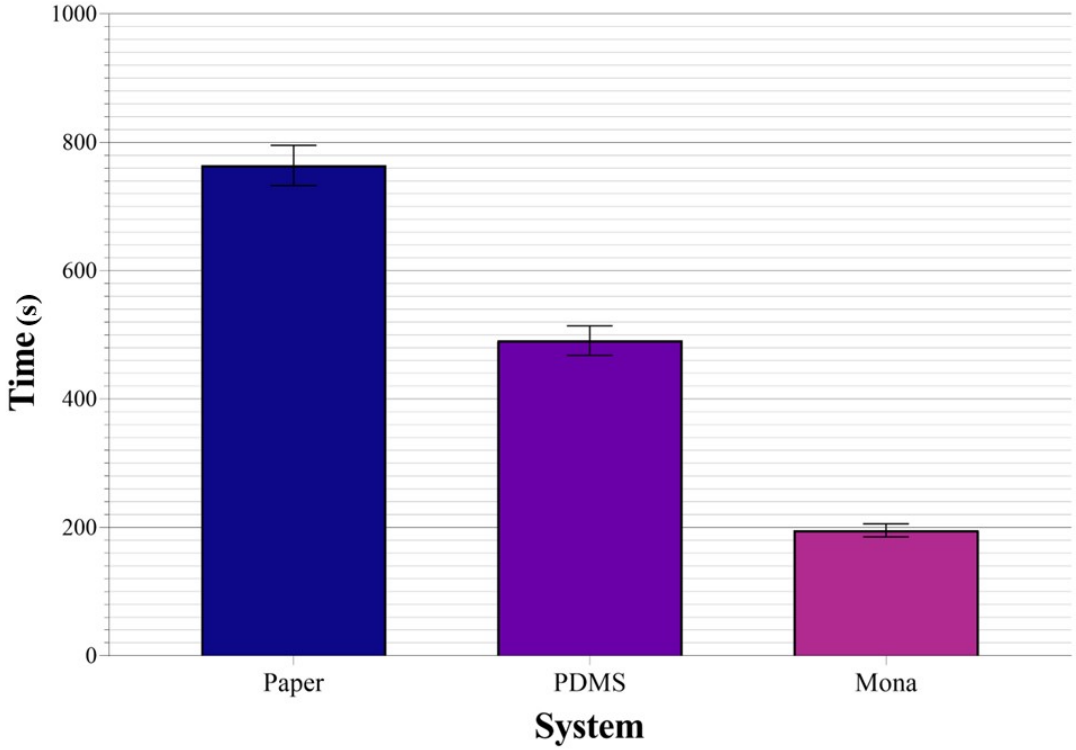
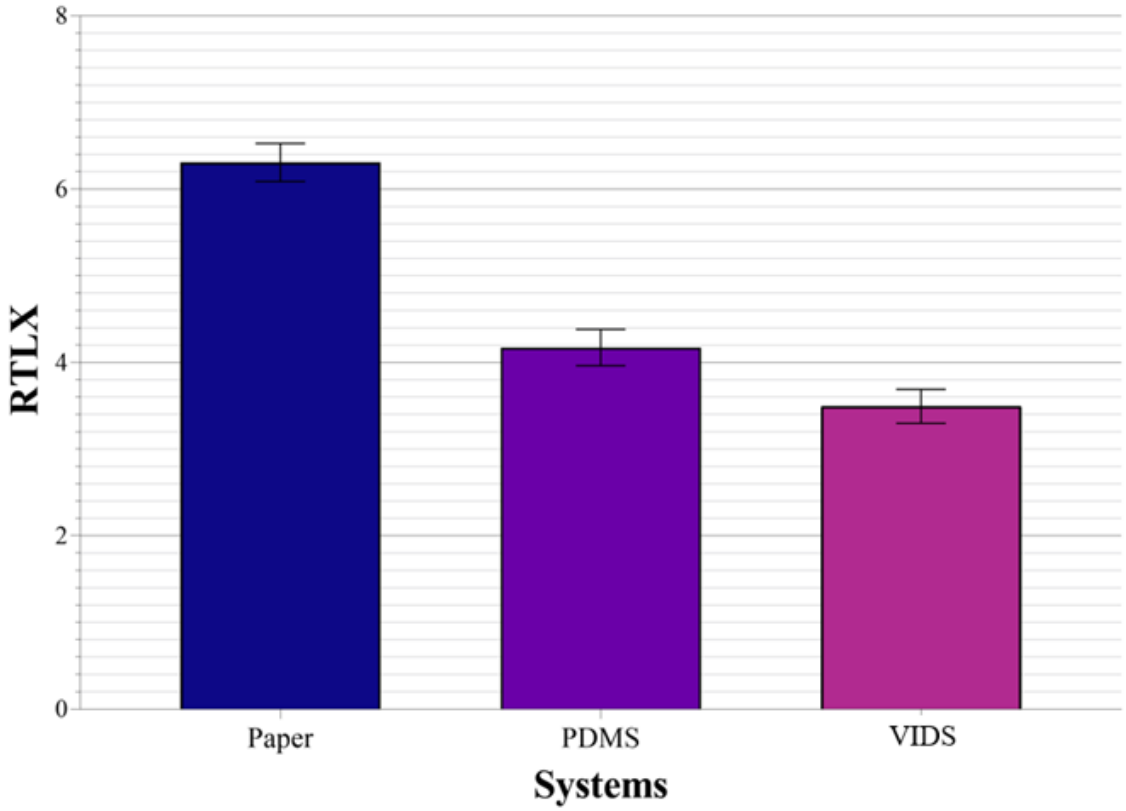


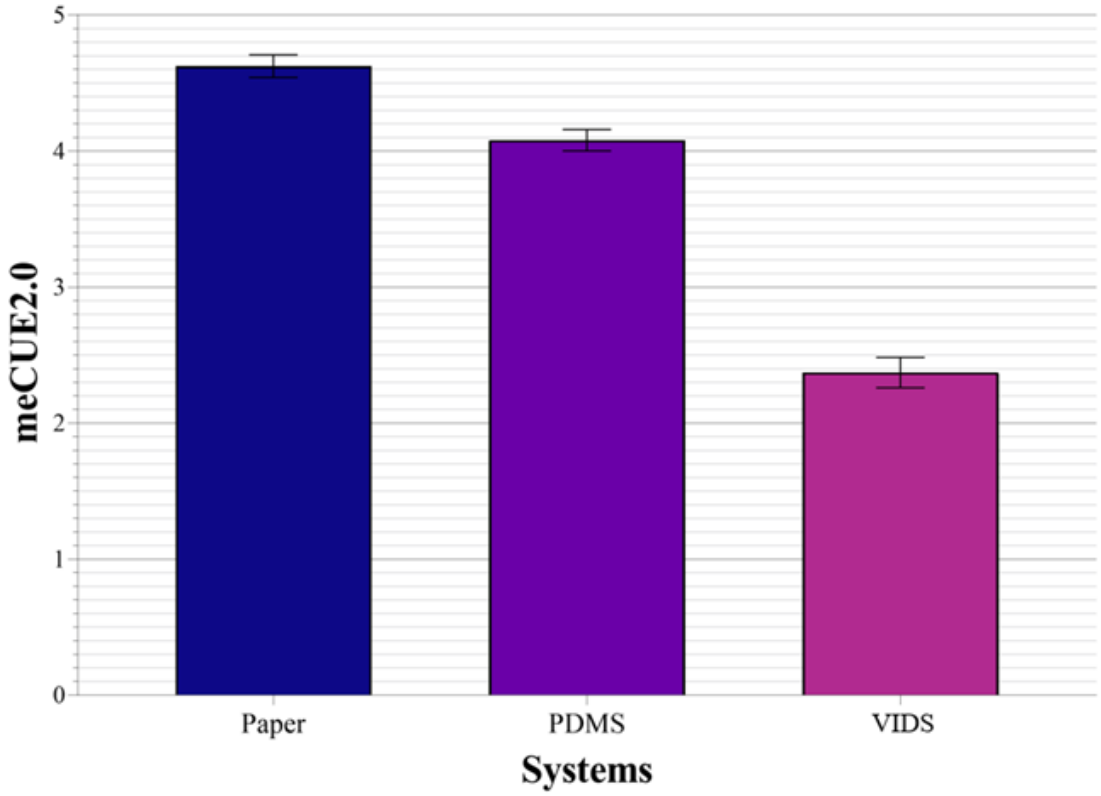
Figure 5. Raw Task Load Index (RTLX) with standard error of the mean. PDMS: patient data management system; VIDS: voice information and documentation system.



Subjective User Experience

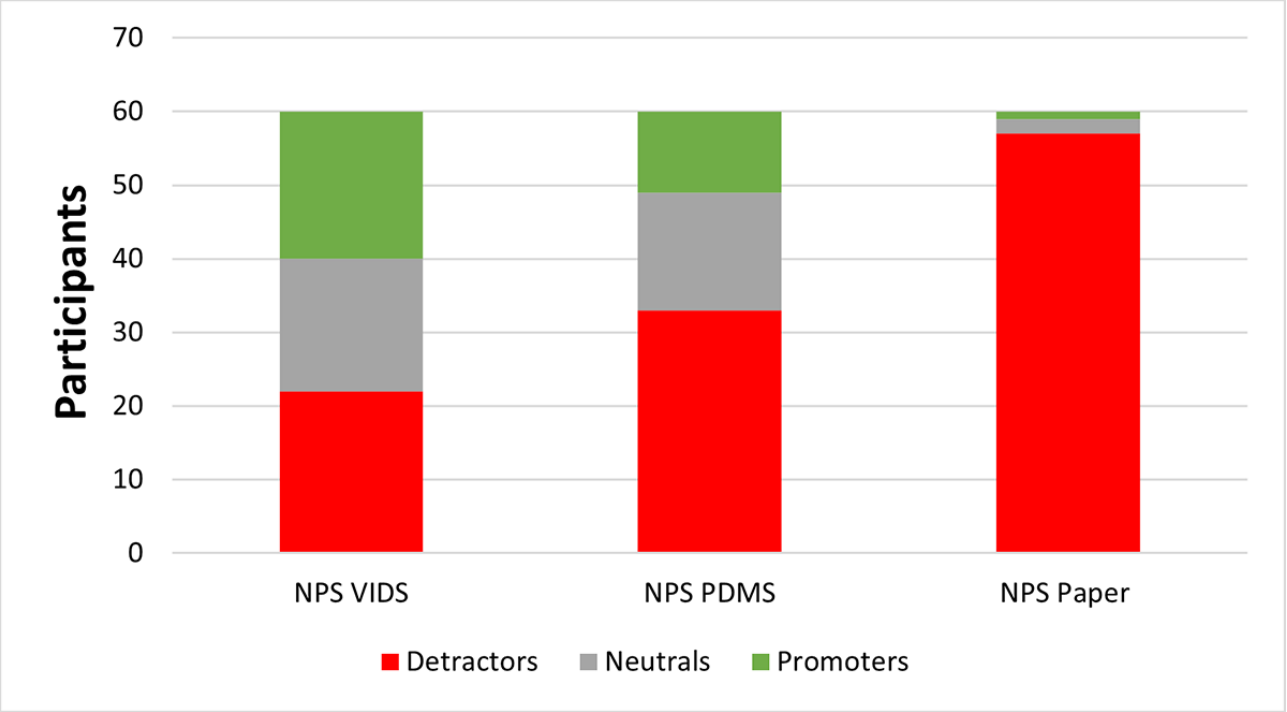
In order to evaluate the subjective user perception, the answers of the meCUE2.0 were analyzed. The average meCUE2.0 score was 2.37 (SD 0.86) for paper documentation, 4.08 (SD 0.61) for PDMS, and 4.63 (SD 0.64) for VIDS (Figure 6). Thus, VIDS showed the highest user satisfaction, while the participants felt least satisfied with using paper-based documentation. The meCUE2.0 was further analyzed using the mean scores of each participant for each of the 3 systems. This assumption was tested using a repeated measures ANOVA followed by pairwise comparisons with Bonferroni correction. The ANOVA showed after Greenhouse-Geisser correction (Mauchly W=0.9; P=.04) a significant difference between the 3 tested systems (n=60; $F_{1.81, 106.78}$=144.73; η_p^2=0.71; P<.001). The effect size is Cohen d=1.56, thus corresponding to a strong effect [48]. The pairwise comparisons with Bonferroni correction proved a statistically significant difference between all 3 systems with regard to user experience (P<.001) [49]. The effect size is Cohen d=0.61 for the comparison of VIDS with PDMS (t_{59}=–4.7), Cohen d=1.87 for VIDS with paper documentation (t_{59}=–14.5), and Cohen d=1.57 for PDMS with paper-based documentation (t_{59}=–12.17). Consequently, the effect size always corresponds to a strong effect [48].

Figure 6. Mean meCUE2.0 score and standard error of the mean. PDMS: patient data management system; VIDS: voice information and documentation system.



For each system, the NPS was calculated according to the given formula (NPS = promotors [% of all participants] minus detractors [% of all participants]) and compared as relative values. The NPS is a score correlating with actual consumer behavior and thus the ability of the company or system to be recommended [46]. The score is divided in promotors (score of 9 or higher), detractors (score of 6 or lower), and neutral users (score of 7-8). The distribution of the promotors and detractors can be seen in Figure 7. The comparison showed that the NPS for VIDS is 11.12 times higher than for PDMS and 28.27 times higher than for paper-based documentation (Figure 7).

Figure 7. Distribution net promoter score (NPS). PDMS: patient data management system; VIDS: voice information and documentation system.



The advantages of new, technologically advanced documentation systems could potentially be based on the technical affinity of the participants. In order to explore this assumption, we analyzed the effect of technical affinity on the results. The participants were grouped into high and low technical affinity groups using a median split according to the results of the TA-EG questionnaire. The analysis showed that the technical affinity has no statistically significant impact on the performance of the systems (Multimedia Appendices 4 and 5). Accordingly, it can be assumed that the prototype is beneficial regardless of the user's affinity for technology.

Discussion

Overview

In reaction to the increasing workload in intensive care medicine and the growing ability to integrate AI applications into clinical routines, new technologies emerge, aiming to improve the treatment of patients with critical illnesses. This study was conducted in order to assess the performance, documentation accuracy, mental workload, and user experience associated with tasks typical of the ICU while using 2 established systems (paper-based documentation [PDMS] and an AI-based [VIDS]).

A total of 60 participants, consisting of physicians, nursing staff, and medical students, completed a set of defined tasks with each system. This was complemented by the completion of several questionnaires. With the time taken, the errors made, and the questionnaires, we compared the 3 systems objectively and based on the users' subjective experiences. The results showed a statistically significant benefit in the objective variables of time and mistakes when using the VIDS. These advantages of the AI-based system are in line with several studies showing that speech recognition can reduce the documentation burden [32,34,35]. The results of this study should encourage the continuation of developing this and other AI-based software to reproduce these results in a real clinical setting.

Furthermore, not only the time needed to complete the tasks but also the errors made compared to the primarily established gold standard were significantly reduced when using the VIDS. This can be explained by the AI-based automated completion of the tasks and answers to the questions asked. Voice recognition in combination with AI reduces the number of necessary steps that the user has to take in order to correctly complete a task or answer a question. Therefore, fewer errors can be made. The significantly reduced error rate might also be due to the data presentation and user interface of the software. As Ahmed et al [17] showed in 2011, the presentation of data has a significant impact on the error rate.

Objective variables are important when evaluating the usefulness of a new system; however, subjective perception and workload are equally relevant. Comparing the RTLX results for each system showed that the VIDS and the PDMS were equivalent, as no statistically significant difference between these 2 systems could be observed. The graphical analysis of the data showed a discernible advantage of the AI-based system; however, this has not proven to be statistically significant. Because we had to limit the required time for study participation, the number of tasks to be completed within the study arms had to be limited as well. The VIDS used in the study was in prototypic stages and has potentially not yet shown its full strengths in terms of usability, consequently reducing mental workload. Therefore, future studies should elaborate on to what extent the further development of the VIDSs contributes to a change in this effect.

The questionnaires aiming to evaluate the subjective user experience confirmed that the participants felt an advantage when using the VIDS compared to PDMS and paper-based documentation (Multimedia Appendix 6). The meCUE 2.0 showed a statistically significantly higher score for the user

experience compared to PDMS and paper documentation. As a high score represents a positive user experience, the results underline that new approaches are not only objectively but also subjectively advantageous in the defined tasks we tested.

These 2 analyses of the subjective user experience were filled out after the participants completed all tasks with every system. The results underline that the VIDS have not only shown equivalence and advantage, respectively, in the objective variables but also in the perceptions of the participants. This goes in line with a 2019 study by Momenipur et al [7] showing that physicians also feel constant time pressure while working and underlined the major importance to improve the subjective work experience and efficiency [7]. The results of the study suggest that by using voice-based software solutions, this can potentially be achieved. This is further underlined by the outcome of the perceived speed of documentation. A total of 80% (48/60) of the participants ranked the VIDS highest. Therefore, the work time was not only statistically significantly less when using the VIDS compared to PDMS and paper-based documentation but also subjectively the lowest. This is especially relevant, as Tajirian et al [50] showed in their study in 2020 that physicians tend to overestimate the time they spend with electronic health records.

Limitations

Clearly, this study has some limitations. High efforts were undertaken in order to reduce confounders and circumstances between the study arms. However, as the 3 observed interventions are very different in technical requirements, input methods, and distribution among ICUs, the study can only be an indication for the acceptance and performance of these systems. Further studies in real-world ICUs are currently in preparation. These studies will analyze the accuracy, efficiency, and mental workload when using the new AI-based VIDS within an operating ICU. These studies will particularly analyze the roots of the observed effects (eg, which technical aspect contributes to what extent) in order to further direct development and progress in this field. Further, the analysis between different VIDSs and different user groups (eg, linguistic backgrounds and age) will be a closely analyzed. Furthermore, the study could not be performed in a real ICU setting due to infection constraints during the COVID-19 pandemic. Even though we tried to reproduce the high noise level of ICUs by performing the tests in a noisy simulation environment, studies in actual ICUs will have to be conducted in order to confirm the results. Another limitation of the study results from the tasks observed. The comparison of 3 different interaction methods, all developed in different decades, using very different technologies, and also requiring different training levels, is challenging and can only give an indication of the value of these systems in real-world health care usage. We attempted to overcome this by using standardization as much as possible and varying interventions in order to limit crossover effects. Additional studies are needed, taking in particular the technical details of VIDS and PDMS (eg, voice recognition, language understanding, data processing, and user interface) into account to make targeted development possible and address potential usability restrictions. As we only tested 3 aspects of the complex work on ICUs with predefined tasks, larger studies will have to confirm the benefits of the new software within the actual workflow of ICUs. Implementing complex IT systems, such as PDMS and VIDS, in a health care workflow certainly produces logistical and economic challenges, as new monitors and systems have to be installed during ongoing patient care in the ICU. Further, networking and IT infrastructure are required, and hospital prerequisites, such as the presence of digital patient data, have to be fulfilled. As documentation is an absolute necessity in patient care, this implementation might lead to disturbances and consequently increased workload, an effect that has to be taken closely into account.

Consequently, in order to assess the full picture of a system's performance, it is required to consider the user's personal perception as well as the objective measurements [50]. The well-being of the medical staff is closely connected to patient safety [10]. In the ranking of the highest user satisfaction, the VIDS was chosen most often on the first rank, as can be seen in Multimedia Appendix 7. This implies that using VIDS can potentially improve the quality of care and patient safety.

Conclusions

A high standardization and objectification of the systems studied was one of the main goals of the study. Nevertheless, the diversity of the investigated systems, the different user interfaces, and the usage contexts inevitably create an inhomogeneity that cannot be completely eliminated. Therefore, the approach of this study was to choose a usage-centric and user-centered object of study. By choosing a diverse set, including novel and well-established metrics, we also tried to focus on the multidimensionality of the results. In order to enable the greatest possible comparability of the developed systems, the adoption of such a measurement method is essential.

Nevertheless, the published approach is only one of the possible solutions to the problem. At this point, we would like to explicitly encourage the use of the protocols we have developed and to further improve and objectify them. In the long term, only based on an established, manufacturer-independent protocol is it possible to approximate the comparison between the different approaches. As novel systems are arising, the proposed study protocol could be the starting point for the development of an industry-wide, vendor-independent accepted standard.

Overall, the results of this study confirm the potential of the use of AI in the clinical setting to reduce workload and improve patient care. The workload in ICUs is growing due to an increasing amount of data collected and the need to document, analyze, and interpret these data points for each patient. In conclusion, AI-based systems like VIDS have the potential to reduce this workload and improve evidence-based and safe patient care.

Conflicts of Interest

AP, LM, and GM are cofounders of Clinomic GmbH. AP and LM are chief executive officers of Clinomic GmbH. GM received restricted research grants and consultancy fees from Braun Melsungen, Biotest, 4Teen4, and Adrenomed outside of the submitted work. AP and LM received consulting fees from Sphingotec GmbH outside of the submitted work. MG is a working student at Clinomic. All other authors declare that they have no conflict of interest.

Multimedia Appendix 1
TA-EG table.
[DOCX File , 13 KB - medinform_v11i1e44773_app1.docx]

Multimedia Appendix 2
meCUE2.0 table.
[DOCX File , 14 KB - medinform_v11i1e44773_app2.docx]

Multimedia Appendix 3
Solution pathways for task completion. Details about solution pathways to all tasks with each system and necessary interaction with each system.
[DOCX File , 17 KB - medinform_v11i1e44773_app3.docx]

Multimedia Appendix 4
Technical affinity table.
[DOCX File , 13 KB - medinform_v11i1e44773_app4.docx]

Multimedia Appendix 5
Correlation technology affinity on performance.
[DOCX File , 107 KB - medinform_v11i1e44773_app5.docx]

Multimedia Appendix 6
User satisfaction table.
[DOCX File , 13 KB - medinform_v11i1e44773_app6.docx]

Multimedia Appendix 7
Working speed table.
[DOCX File , 52 KB - medinform_v11i1e44773_app7.docx]

References

1. Mayrhofer O. Definition, funktion und bedeutung der intensivmedizin. In: Frey R, Mayrhofer O, Hügin W, editors. Lehrbuch der Anaesthesiologie, Reanimation und Intensivtherapie. Berlin Heidelberg: Springer; 1972:881-883.
2. Glas M, Pfortmüller C. Mein Erster Dienst—Intensivmedizin. Berlin Heidelberg: Springer; 2020.
3. Flohr L, Beaudry S, Johnson KT, West N, Burns CM, Ansermino JM, et al. Clinician-driven design of VitalPAD-an intelligent monitoring and communication device to improve patient safety in the intensive care unit. IEEE J Transl Eng Health Med 2018;6:3000114 [FREE Full text] [doi: 10.1109/JTEHM.2018.2812162] [Medline: 29552425]
4. Butler R, Monsalve M, Thomas GW, Herman T, Segre AM, Polgreen PM, et al. Estimating time physicians and other health care workers spend with patients in an intensive care unit using a sensor network. Am J Med 2018;131(8):972.e9-972.e15. [doi: 10.1016/j.amjmed.2018.03.015] [Medline: 29649458]
5. Hefter Y, Madahar P, Eisen LA, Gong MN. A time-motion study of ICU workflow and the impact of strain. Crit Care Med 2016;44(8):1482-1489. [doi: 10.1097/CCM.0000000000001719] [Medline: 27058466]
6. Carayon P, Wetterneck TB, Alyousef B, Brown RL, Cartmill RS, McGuire K, et al. Impact of electronic health record technology on the work and workflow of physicians in the intensive care unit. Int J Med Inform 2015;84(8):578-594 [FREE Full text] [doi: 10.1016/j.ijmedinf.2015.04.002] [Medline: 25910685]
7. Momenipur A, Pennathur PR. Balancing documentation and direct patient care activities: a study of a mature electronic health record system. Int J Ind Ergon 2019;72:338-346 [FREE Full text] [doi: 10.1016/j.ergon.2019.06.012] [Medline: 32201437]
8. Wright AA, Katz IT. Beyond burnout—redesigning care to restore meaning and sanity for physicians. N Engl J Med 2018;378(4):309-311. [doi: 10.1056/NEJMp1716845] [Medline: 29365301]

9. Sinsky C, Colligan L, Li L, Prgomet M, Reynolds S, Goeders L, et al. Allocation of physician time in ambulatory practice: a time and motion study in 4 specialties. Ann Intern Med 2016;165(11):753-760 [FREE Full text] [doi: 10.7326/M16-0961] [Medline: 27595430]
10. Hall LH, Johnson J, Watt I, Tsipa A, O'Connor DB. Healthcare staff wellbeing, burnout, and patient safety: a systematic review. PLoS One 2016;11(7):e0159015 [FREE Full text] [doi: 10.1371/journal.pone.0159015] [Medline: 27391946]
11. Collins S, Couture B, Kang MJ, Dykes P, Schnock K, Knaplund C, et al. Quantifying and visualizing nursing flowsheet documentation burden in acute and critical care. AMIA Annu Symp Proc 2018;2018:348-357 [FREE Full text] [Medline: 30815074]
12. Fischer JE, Calame A, Dettling AC, Zeier H, Fanconi S. Experience and endocrine stress responses in neonatal and pediatric critical care nurses and physicians. Crit Care Med 2000;28(9):3281-3288. [doi: 10.1097/00003246-200009000-00027] [Medline: 11008993]
13. Densen P. Challenges and opportunities facing medical education. Trans Am Clin Climatol Assoc 2011;122:48-58 [FREE Full text] [Medline: 21686208]
14. Beecken WD. Changes—Analyse der Entwicklung der Digitalen Medizin im deutschen Gesundheitssystem aus ärztlicher Sicht. In: Matusiewicz D, editor. Think Tanks im Gesundheitswesen: Deutsche Denkfabriken und ihre Positionen zur Zukunft der Gesundheit. Germany: Springer Fachmedien Wiesbaden; 2020:29-44.
15. Martin L, Peine A. [What is new… implementation of artificial intelligence in intensive care medicine: hype or already reality?]. Anaesthesist 2021;70(1):40-41. [doi: 10.1007/s00101-020-00891-7] [Medline: 33242115]
16. Chao CA. The impact of electronic health records on collaborative work routines: a narrative network analysis. Int J Med Inform 2016;94:100-111. [doi: 10.1016/j.ijmedinf.2016.06.019] [Medline: 27573317]
17. Ahmed A, Chandra S, Herasevich V, Gajic O, Pickering BW. The effect of two different electronic health record user interfaces on intensive care provider task load, errors of cognition, and performance. Crit Care Med 2011;39(7):1626-1634. [doi: 10.1097/CCM.0b013e31821858a0] [Medline: 21478739]
18. Toll E. The cost of technology. JAMA 2020;323(17):1661-1662. [doi: 10.1001/jama.2020.2752] [Medline: 32369125]
19. Shanafelt TD, West CP, Sloan JA, Novotny PJ, Poland GA, Menaker R, et al. Career fit and burnout among academic faculty. Arch Intern Med 2009;169(10):990-995 [FREE Full text] [doi: 10.1001/archinternmed.2009.70] [Medline: 19468093]
20. Grol R, Mokkink H, Smits A, van Eijk J, Beek M, Mesker P, et al. Work satisfaction of general practitioners and the quality of patient care. Fam Pract 1985;2(3):128-135. [doi: 10.1093/fampra/2.3.128] [Medline: 4043602]
21. Manomenidis G, Panagopoulou E, Montgomery A. Job burnout reduces hand hygiene compliance among nursing staff. J Patient Saf 2019;15(4):e70-e73. [doi: 10.1097/PTS.0000000000000435] [Medline: 29028691]
22. Muinga N, Abejirinde IOO, Paton C, English M, Zweekhorst M. Designing paper-based records to improve the quality of nursing documentation in hospitals: a scoping review. J Clin Nurs 2021;30(1-2):56-71 [FREE Full text] [doi: 10.1111/jocn.15545] [Medline: 33113237]
23. Fröhlich D, Bittersohl C, Schroeder K, Schöttle D, Kowalinski E, Borgwardt S, et al. Reliability of paper-based routine documentation in psychiatric inpatient care and recommendations for further improvement. Front Psychiatry 2019;10:954 [FREE Full text] [doi: 10.3389/fpsyt.2019.00954] [Medline: 32009991]
24. Castellanos I, Ganslandt T, Prokosch HU, Schüttler J, Bürkle T. [Implementation of a patient data management system. Effects on intensive care documentation]. Anaesthesist 2013;62(11):887-897 [FREE Full text] [doi: 10.1007/s00101-013-2239-x] [Medline: 24126951]
25. Ballermann MA, Shaw NT, Arbeau KJ, Mayes DC, Gibney RTN. Impact of a critical care clinical information system on interruption rates during intensive care nurse and physician documentation tasks. Stud Health Technol Inform 2010;160(Pt 1):274-278. [Medline: 20841692]
26. Cheung A, van Velden FHP, Lagerburg V, Minderman N. The organizational and clinical impact of integrating bedside equipment to an information system: a systematic literature review of Patient Data Management Systems (PDMS). Int J Med Inform 2015;84(3):155-165. [doi: 10.1016/j.ijmedinf.2014.12.002] [Medline: 25601332]
27. Baumann LA, Baker J, Elshaug AG. The impact of electronic health record systems on clinical documentation times: a systematic review. Health Policy 2018;122(8):827-836. [doi: 10.1016/j.healthpol.2018.05.014] [Medline: 29895467]
28. Hodgson T, Magrabi F, Coiera E. Efficiency and safety of speech recognition for documentation in the electronic health record. J Am Med Inform Assoc 2017;24(6):1127-1133 [FREE Full text] [doi: 10.1093/jamia/ocx073] [Medline: 29016971]
29. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. Nat Med 2019;25(1):24-29. [doi: 10.1038/s41591-018-0316-z] [Medline: 30617335]
30. Dymek C, Kim B, Melton GB, Payne TH, Singh H, Hsiao CJ. Building the evidence-base to reduce electronic health record-related clinician burden. J Am Med Inform Assoc 2021;28(5):1057-1061 [FREE Full text] [doi: 10.1093/jamia/ocaa238] [Medline: 33340326]
31. Hakes B, Whittington J. Assessing the impact of an electronic medical record on nurse documentation time. Comput Inform Nurs 2008;26(4):234-241. [doi: 10.1097/01.NCN.0000304801.00628.ab] [Medline: 18600132]
32. Goss FR, Blackley SV, Ortega CA, Kowalski LT, Landman AB, Lin CT, et al. A clinician survey of using speech recognition for clinical documentation in the electronic health record. Int J Med Inform 2019;130:103938. [doi: 10.1016/j.ijmedinf.2019.07.017] [Medline: 31442847]

33. Zuchowski M. Medizinische spracherkennung: weniger dokumentationsaufwand, mehr zeit: seit einigen jahren wird die medizinische spracherkennung regelmäßig zur unterstützung der dokumentation in krankenhäusern eingesetzt. Dies beeinflusst die interdisziplinäre zusammenarbeit und spart zeit, die stattdessen in patientennahe tätigkeiten einfließen kann. Ihr volles Potenzial kann medizinische Spracherkennung allerdings erst entfalten, wenn sie mit anderen Anwendungen vernetzt ist. KMA 2020;25(04):72-73. [doi: 10.1055/s-0040-1709881]
34. Blackley SV, Schubert VD, Goss FR, Al Assad W, Garabedian PM, Zhou L. Physician use of speech recognition versus typing in clinical documentation: a controlled observational study. Int J Med Inform 2020;141:104178. [doi: 10.1016/j.ijmedinf.2020.104178] [Medline: 32521449]
35. Hodgson T, Coiera E. Risks and benefits of speech recognition for clinical documentation: a systematic review. J Am Med Inform Assoc 2016;23(e1):e169-e179 [FREE Full text] [doi: 10.1093/jamia/ocv152] [Medline: 26578226]
36. Licht A, Blaser J. [Speech recognition in clinical routine, a pilot trial at the Zurich University Hospital]. Praxis (Bern 1994) 2002;91(19):831-835. [doi: 10.1024/0369-8394.91.19.831] [Medline: 12071083]
37. Poncette AS, Spies C, Mosch L, Schieler M, Weber-Carstens S, Krampe H, et al. Clinical requirements of future patient monitoring in the intensive care unit: qualitative study. JMIR Med Inform 2019;7(2):e13064 [FREE Full text] [doi: 10.2196/13064] [Medline: 31038467]
38. Cosgriff CV, Celi LA, Stone DJ. Critical care, critical data. Biomed Eng Comput Biol 2019;10:1179597219856564 [FREE Full text] [doi: 10.1177/1179597219856564] [Medline: 31217702]
39. IntelliSpace critical care and anesthesia: release H technical data sheet. Koninklijke Philips N.V. 2015. URL: https://images.philips.com/is/content/PhilipsConsumer/Campaigns/HC20140401_DG/Documents/ICCA%20data%20sheet.pdf?_ga=2.6228528.2006153948.1633410240-1425734618.1633410240 [accessed 2021-10-05]
40. Consortium Uniklinik RWTH Aachen, RWTH Aachen, SERMAS, CapDigital, ATOS, Hochschule Trier. DEL05 final report: 20655—clinical artificial intelligence improving healthcare. Claire, The Virtual Healthcare Assistant. 2021. URL: https://www.umwelt-campus.de/en/forschung/projekte/projekte-entdecken/claire [accessed 2023-10-19]
41. Karrer K, Glaser C, Clemens C, Bruder C. Berliner werkstatt mensch-maschine-systeme. Beiträge 8. 2009. URL: https://idw-online.de/en/event?print=1&id=28067 [accessed 2023-10-19]
42. Hart SG. Nasa-Task Load Index (NASA-TLX); 20 years later. Proc Hum Factors Ergon Soc Annu Meet 2006;50(9):904-908 [FREE Full text] [doi: 10.1177/154193120605000909]
43. Hendy KC, Hamilton KM, Landry LN. Measuring subjective workload: when is one scale better than many? Hum Factors 1993;35(4):579-601 [FREE Full text] [doi: 10.1177/001872089303500401]
44. Mital A. Advances in Industrial Ergonomics and Safety: v. 1 (Proceedings of the Annual International Industrial Ergonomics & Safety Conference held in Cincinnati, Ohio, USA, 5-9 June 1989). Oxfordshire: Taylor & Francis Ltd; 1989.
45. Minge M. Nutzererleben messen mit dem meCUE 2.0—Ein Tool für alle Fälle? 2018 Presented at: Mensch und Computer 2018—Workshopband; Dresden; September 2-5, 2018.
46. Reichheld FF. The one number you need to grow. Harv Bus Rev 2003;81(12):46-54, 124. [Medline: 14712543]
47. Eid M, Gollwitzer M, Schmitt M. Statistik und Forschungsmethoden: Lehrbuch. Mit Online-Material. Weinheim Basel: Beltz Verlag; 2015.
48. Cohen J. Statistical Power Analysis for the Behavioral Sciences. Hoboken: Taylor and Francis; 2013.
49. Bland JM, Altman DG. Multiple significance tests: the Bonferroni method. BMJ 1995;310(6973):170 [FREE Full text] [doi: 10.1136/bmj.310.6973.170] [Medline: 7833759]
50. Tajirian T, Stergiopoulos V, Strudwick G, Sequeira L, Sanches M, Kemp J, et al. The influence of electronic health record use on physician burnout: cross-sectional survey. J Med Internet Res 2020;22(7):e19274 [FREE Full text] [doi: 10.2196/19274] [Medline: 32673234]

Abbreviations

AI: artificial intelligence
ICU: intensive care unit
NASA-TLX: National Aeronautics and Space Administration Task Load Index
NPS: net promoter score
PDMS: patient data management system
RTLX: Raw Task Load Index
TA-EG: Fragebogen zur Erfassung der Technikaffinität als Umgang mit und Einstellung zuelektronischen Geräte (“Questionnaire for the assessment of technology affinity as handling and attitude toward electronic devices”)
VIDS: voice information and documentation system

Edited by Q Chen; submitted 02.12.22; peer-reviewed by C Bérubé, D Barra, E Toki; comments to author 01.03.23; revised version received 21.06.23; accepted 17.10.23; published 28.11.23.

Please cite as:
Peine A, Gronholz M, Seidl-Rathkopf K, Wolfram T, Hallawa A, Reitz A, Celi LA, Marx G, Martin L
Standardized Comparison of Voice-Based Information and Documentation Systems to Established Systems in Intensive Care: Crossover Study
JMIR Med Inform 2023;11:e44773
URL: https://medinform.jmir.org/2023/1/e44773
doi:10.2196/44773
PMID:38015593

Original Paper

A Comprehensive and Improved Definition for Hospital-Acquired Pressure Injury Classification Based on Electronic Health Records: Comparative Study

Mani Sotoodeh[1], BSc, MSc, PhD; Wenhui Zhang[2], BSN, MSc, PhD; Roy L Simpson[2], DNP, RN, DPNAP; Vicki Stover Hertzberg[2], PhD, PStat; Joyce C Ho[3], BSc, MSc, PhD

[1]Public Health Research Institute of University of Montreal, University of Montreal, Montreal, QC, Canada

[2]School of Nursing, Emory University, Atlanta, GA, United States

[3]Department of Computer Science, Emory University, Atlanta, GA, United States

Corresponding Author:
Wenhui Zhang, BSN, MSc, PhD
School of Nursing
Emory University
Room 224
1520 Clifton Rd
Atlanta, GA, 30322
United States
Phone: 1 512 968 8985
Email: wenhui.zhang@emory.edu

Abstract

Background: Patients develop pressure injuries (PIs) in the hospital owing to low mobility, exposure to localized pressure, circulatory conditions, and other predisposing factors. Over 2.5 million Americans develop PIs annually. The Center for Medicare and Medicaid considers hospital-acquired PIs (HAPIs) as the most frequent preventable event, and they are the second most common claim in lawsuits. With the growing use of electronic health records (EHRs) in hospitals, an opportunity exists to build machine learning models to identify and predict HAPI rather than relying on occasional manual assessments by human experts. However, accurate computational models rely on high-quality HAPI data labels. Unfortunately, the different data sources within EHRs can provide conflicting information on HAPI occurrence in the same patient. Furthermore, the existing definitions of HAPI disagree with each other, even within the same patient population. The inconsistent criteria make it impossible to benchmark machine learning methods to predict HAPI.

Objective: The objective of this project was threefold. We aimed to identify discrepancies in HAPI sources within EHRs, to develop a comprehensive definition for HAPI classification using data from all EHR sources, and to illustrate the importance of an improved HAPI definition.

Methods: We assessed the congruence among HAPI occurrences documented in clinical notes, diagnosis codes, procedure codes, and chart events from the Medical Information Mart for Intensive Care III database. We analyzed the criteria used for the 3 existing HAPI definitions and their adherence to the regulatory guidelines. We proposed the Emory HAPI (*EHAPI*), which is an improved and more comprehensive HAPI definition. We then evaluated the importance of the labels in training a HAPI classification model using tree-based and sequential neural network classifiers.

Results: We illustrate the complexity of defining HAPI, with <13% of hospital stays having at least 3 PI indications documented across 4 data sources. Although chart events were the most common indicator, it was the only PI documentation for >49% of the stays. We demonstrate a lack of congruence across existing HAPI definitions and *EHAPI*, with only 219 stays having a consensus positive label. Our analysis highlights the importance of our improved HAPI definition, with classifiers trained using our labels outperforming others on a small manually labeled set from nurse annotators and a consensus set in which all definitions agreed on the label.

Conclusions: Standardized HAPI definitions are important for accurately assessing HAPI nursing quality metric and determining HAPI incidence for preventive measures. We demonstrate the complexity of defining an occurrence of HAPI, given the conflicting and incomplete EHR data. Our *EHAPI* definition has favorable properties, making it a suitable candidate for HAPI classification tasks.

(JMIR Med Inform 2023;11:e40672) doi:10.2196/40672

KEYWORDS

pressure ulcer; decubitus ulcer; electronic medical records; bedsore; nursing; data mining; electronic health record; EHR; nursing assessment; pressure ulcer care; pressure ulcer prevention; EHR data; EHR systems; nursing quality

Introduction

Background and Significance

Hospital-Acquired Pressure Injury, a Key Nursing Metric

Localized damage to the skin or underlying tissues characterizes pressure injury (PI). PI is typically found over a bony prominence or under a medical device and can be caused by lying down or sitting in one place for too long without much movement [1,2]. Hospital-acquired PI (HAPI) is classified according to the PI stage and the time of its development or progression. HAPI is associated with extended hospital stays, high readmission rates, reduced quality of life, and mortality [3]. HAPI is the most frequent preventable adverse event in hospitals according to the Center for Medicare and Medicaid (CMS) and the second most common claim in wrongful death lawsuits [3]. CMS and the Agency for Healthcare Research and Quality (AHRQ) consider HAPI a "never event," that is, events with profound financial penalties to providers on reimbursement [4]. More than 2000 US hospitals are part of the National Database of Nursing Quality Indicators program to measure nursing quality metrics, that is, events that are directly associated with the quality of nursing care. The National Database of Nursing Quality Indicators requires participating facilities to perform a quarterly survey of patients to estimate the incidence of HAPI [2]. Thus, accurate information on the incidence of HAPI in a health care unit is critical for assessing nursing quality and planning by hospital administrators.

Electronic Health Records and HAPI Identification, Opportunities, and Challenges

Electronic health records (EHRs) provide extensive information on existing and new PIs, including diagnosis codes; characteristics in structured charts, such as stage, depth, and location of PIs; and PI keywords in semistructured or unstructured clinical notes. Automatic detection of HAPI in EHRs using computational models facilitates clinical decision-making and patient care [5]. Predictive models for HAPI depend on the quality, dependability, and consistency of the data set. However, the complexity and subjectivity of PI screening, detection, and staging impact the reliability of PI documentation. PI documentation reliability also depends on the competency and continuity of nursing staff and their roles as well as changes in data entry or the EHR system.

Despite advances in prevention and treatment, HAPI persists and is difficult to identify from EHRs. Data sources provide contradictory information on PIs. Furthermore, the predictive model accuracy relies heavily on the definition of HAPI and accurate labels. Previous studies using EHR data have used inconsistent definitions of the HAPI. Some describe medical conditions that indicate HAPI [2,6,7]; some identify HAPI in all records associated with a hospital stay [7-9]; and others use prior laboratory data to predict HAPI [10].

Inconsistent HAPI labels adversely impact the model performance in HAPI classification and complicate comparison of multiple models. Thus, the correct identification of HAPI labels from EHR data is essential for HAPI studies. Assessing the performance of machine learning models for HAPI tasks using fixed benchmark data requires accessing appropriate clinical data from EHR databases, unifying multiple data sources, and using them consistently with regulatory guidelines. Here, we propose a HAPI definition that meets these requirements.

Toward a Unified HAPI Definition and More Accurate HAPI Classification

We illustrate the challenges in detecting HAPI in EHRs using the Medical Information Mart for Intensive Care III (MIMIC-III) [11] as a case study. MIMIC-III is one of the most widely used open benchmark data sets, built over CareVue and Metavision EHR systems that encompass approximately 59,000 hospital stays. The patient data include demographics, vital signs, laboratory results, physiological measurements, diagnoses, and clinical and nursing notes.

This study highlights the gaps between existing HAPI definitions for MIMIC-III and the guidelines set by CMS and other regulatory bodies. We propose the Emory HAPI (*EHAPI*) definition, which better adheres to the regulatory guidelines. We then demonstrate the impact of our improved definition in training a more accurate HAPI classification model. The classification performance was evaluated using a manually labeled set from nurse annotators as a proxy for the HAPI ground truth.

Our main contributions are as follows:

1. An improved HAPI definition that leverages diverse data sources and accounts for their reliability while adhering more closely to clinical guidelines
2. Illustrating the impact of the noncomprehensive HAPI definition on training a HAPI prediction model

To achieve these objectives, we pursue the following steps: (1) describe challenges in finding evidence for HAPI within different sources in the MIMIC-III data set, (2) use nursing expertise with clinical information to prioritize and combine conflicting data sources, (3) establish core parameters for a practically reasonable HAPI definition, and (4) determine the impact of the definition on the performance of tree-based and neural network–based HAPI classifiers.

Methods

Overview of Data Sources for PI in Hospital Stays

There are 4 major sources of EHR data that may contain information on PI: patient chart events, diagnosis codes, notes, and procedures performed.

Chart Events

Chart events constitute the largest portion of structured clinical data and include many medical services, including laboratory tests, vital signs, nurses' assessments, and general indicators such as patient mental status. Chart events are time-stamped and provide information on the time and order of events during the hospital stay.

Diagnosis Codes

For billing purposes, each hospital stay contains a limited set of diagnosis codes. These codes usually include the most important diagnoses during the hospital stay; however, financial concerns and imperfect mapping of clinical findings to predetermined codes can impact this.

Notes

Clinical notes include any unstructured text information such as radiography reports, electrocardiogram reports, discharge summaries, admission notes, and daily notes made by the care team.

Procedure Codes

Procedure codes indicate timed medical services and surgeries.

PI staging is a core element of the HAPI deterioration status from admission to discharge. Most modern clinical information systems, including the Metavision and CareVue clinical information systems in the MIMIC-III and Emory Healthcare's clinical data warehouse, contain PI staging events and notes. Multimedia Appendix 1 [12] summarizes the details of the PI data sources in the MIMIC-III.

Ideal HAPI Criteria Based on Guidelines

Regulatory authorities identify HAPI using many elements, including the presence of PIs at admission and discharge, changes in stages, unit transfers during admission, and patient death. CMS provides several inclusion and exclusion criteria for HAPI [13-15]. One inclusion criterion is the presence of one or more new or worsened PIs at discharge compared with admission. This includes stages 2 to 4, or PIs not staged owing to slough or eschar, nonremovable dressing or device, or deep tissue injury. Another inclusion criterion is an unstageable PI on admission that is later staged. This is coded on discharge assessment as "present on admission," with the earliest assessed numerical stage. A patient stay is excluded if data on new or worsened stages 2, 3, and 4, or unstageable pressure ulcers, including deep tissue injuries, are missing on the planned or unplanned discharge assessment. In addition, a patient stay is excluded when the patient died during the hospital stay.

The standard practice for newly admitted patients is the completion of admission assessment, as close as possible to the time of admission and within 24 hours. AHRQ also suggests "performance of comprehensive skin assessment within 24 hours of admission" to accurately assess PI rates [16]. The National Pressure Injury Advisory Panel (NPIAP) reference guide [17] defines the facility-acquired rate as the "percentage of individuals who did not have a pressure injury on admission who acquire a pressure injury during their stay in the facility."

Existing MIMIC-III HAPI Case Definitions and Their Limitations

There are 4 existing HAPI definitions for MIMIC-III, which are summarized in the subsequent section. Detailed flowcharts for the various definitions are provided in Multimedia Appendix 1.

Recurrent additive network for temporal risk prediction (CANTRIP) [10] focused on predicting HAPI 48 to 96 hours before its first appearance, or date of event (DOE). The DOE was defined as the first occurrence of either mention of PI-related keywords in time-stamped hospital notes or a PI staging chart event (≥stage 1) >48 hours after admission. Other stays without a DOE were marked as controls. Unfortunately, the CANTRIP case definition included deceased patients and healed or improved PIs.

Cramer et al [6] sought to develop a screening tool for PI by using the first 24 hours of data. They identified HAPI cases using only the PI staging chart events occurring 24 hours after admission. It excluded stage 1 PIs and "unable to stage" and deep tissue injury PIs. Similar to CANTRIP, the Cramer case definition included deceased patients and healed or improved PIs. Other stays constituted the control group.

Sotoodeh et al [9] explored the use of negation preprocessing on clinical text to detect PI. Case patients were defined using International Classification of Diseases (ICD)-9 codes or PI-specific keywords in the clinical notes. Similar to CANTRIP and Cramer definitions, deceased, healed, or improved PIs were included in the case definition. However, in contrast to the CANTRIP and Cramer definitions, they did not consider PI staging chart events. Control stays were defined as the absence of both ICD-9 codes and PI-specific keywords.

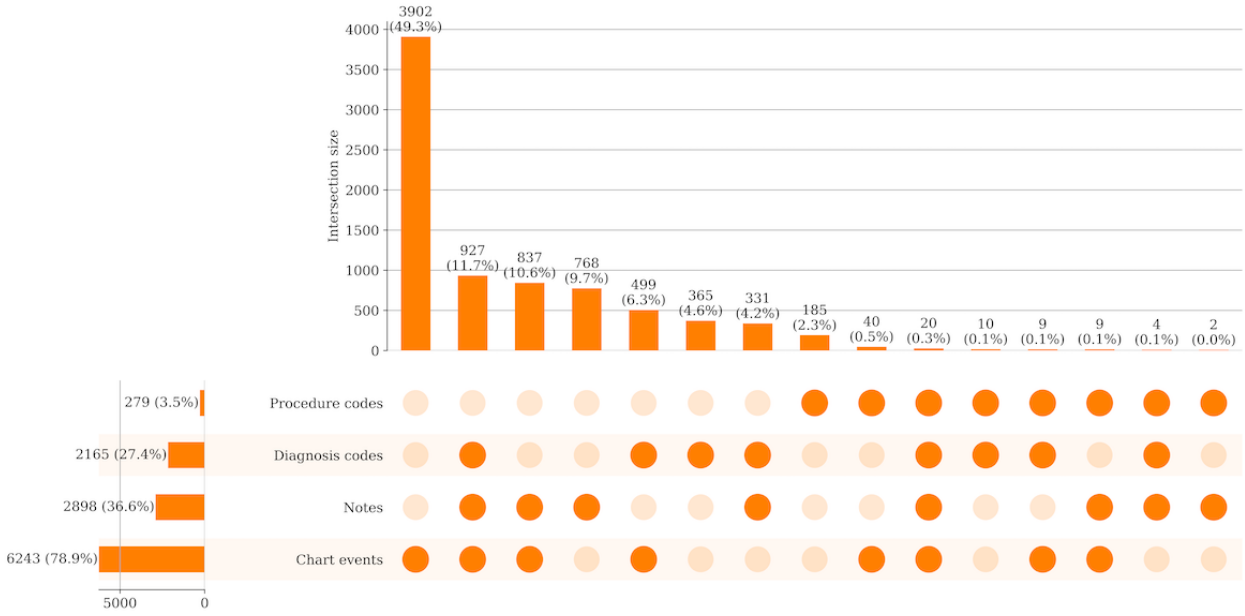
Cox et al [7] focused on identifying appropriate risk factors for PI by using selected variables from the existing literature. They identified a subset of patients who did not have preexisting PI on admission. However, the inclusion and exclusion criteria for identifying HAPI were not explicitly mentioned and are therefore not presented here.

Other studies have focused on predicting HAPI by using other EHR databases. Ranzani et al [8] focused on predicting PI within 30 days of intensive care unit admission in the first 24 hours. They excluded patients who had a preexisting PI on admission or developed PI within the first 48 hours. The case definition was similar to that of CANTRIP, except that notes were not used. Song et al [18] also proposed an early assessment tool for PI risk using 28 relevant features from existing literature. However, the case definition was not discussed in detail. Finally, Hyun et al [19] developed a machine learning model to predict the HAPI. HAPI cases were defined as those containing an ICD-9 code associated with a PI.

EHAPI Case Definition in MIMIC-III

On the basis of existing and ideal HAPI criteria, we identified several essential elements to create a HAPI case definition using EHR data and applied it to MIMIC-III. MIMIC-III has limitations, that is, incongruence of data sources regarding PIs presence and complexity of extracting stage data to verify PI deterioration criteria from admission to discharge, for stays with only comments about PIs in nursing notes and not as timed structured data (Figure 1 and section *Limitations and Future Work*). These limitations inform this exemplar MIMIC-III HAPI case definition.

Figure 1. Medical Information Mart for Intensive Care III (MIMIC-III) data sources consistency for pressure injury (PI) hospital stays.



For the HAPI criteria, we can include or exclude deceased patients, set a minimum age, and consider either 24 or 48 hours from admission to determine admission PIs status. Further decisions for HAPI criteria include the set of clinical events related to PI staging data, the minimum numerical stage for HAPI, numerical stage values assigned to deep tissue injury and unstageable PIs, and the inclusion or exclusion of healed or improved PIs at discharge. Moreover, in addition to staging events, to determine HAPI labels, we considered the presence of certain keywords or diagnosis codes in the notes. We propose a more comprehensive version of the previous definitions, *EHAPI*. The *EHAPI* definition is based on the updated version of the HAPI criteria as determined by the CMS, NPIAP, and AHRQ guidelines [2,14-16]. We extracted data from the admission, patient, and intensive care unit stay tables in the MIMIC-III to construct features and remove irrelevant stays from our analysis.

Our case definition considered only new PIs or PIs that deteriorated by discharge, which required determining the PI stages at admission and discharge for each hospital stay. If a patient had multiple hospital stays, we treated each stay separately. Moreover, in MIMIC-III, a hospital stay encompasses ≥1 intensive care unit stays. Staging occurred within 24 hours of admission. Stage 4 is deep PI, and "unable to stage" was coded as 0. In the absence of the PI stage information at admission, the stage was set to 0. Discharge stage was set as the last recorded stage above 2 occurring later than 24 hours of admission, considering deep tissue injury as stage 3 and "unable to stage" as stage 5. "Unable to stage" was set to stage 5 to capture all possible HAPI irrespective of the admission stage. On the basis of NPIAP documentation [20], deep tissue injury was either stage 3 or 4 PI. Therefore, to allow exclusion from the HAPI criteria because of stage improvement during the stay, we coded deep tissue injury as stage 4 at admission and stage 3 at discharge.

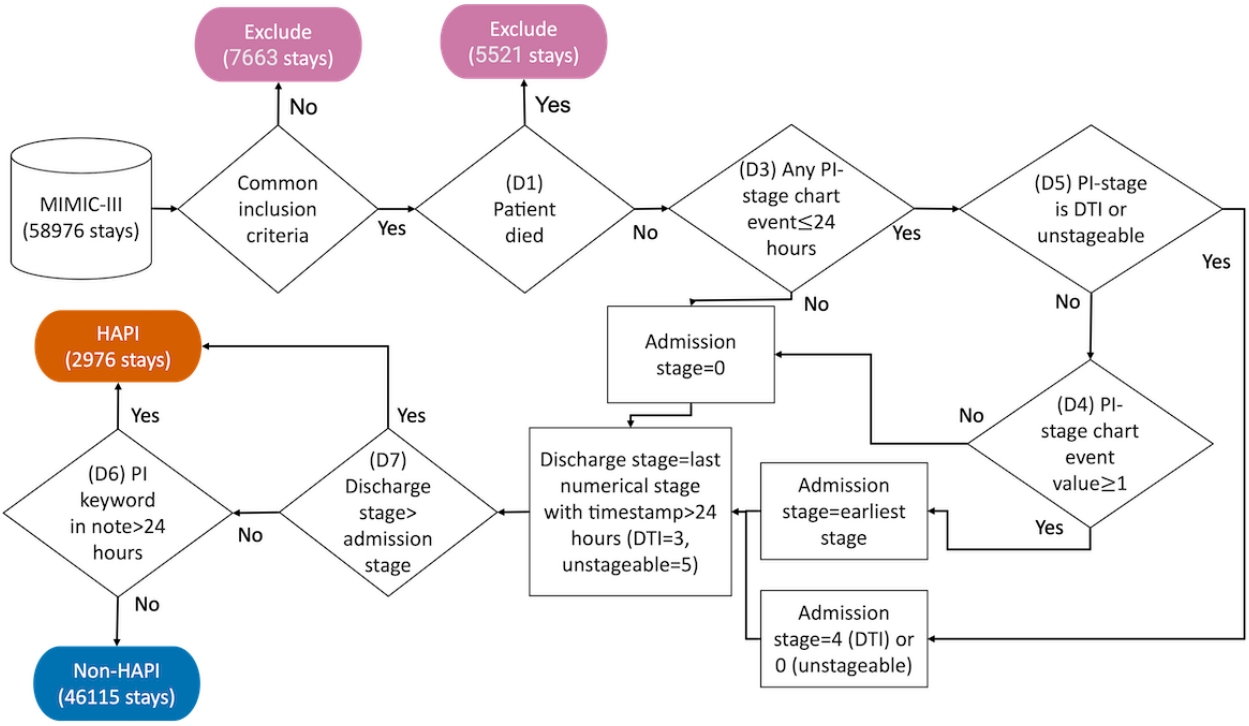
We excluded the stays that did not meet the common inclusion criteria. The common inclusion criteria across the four definitions were as follows: (1) presence of at least one clinical note, (2) documented discharge time as after admission time, (3) the patients aged <15 years, and (4) no admission documentation of a PI. *EHAPI* excluded patients who died in the hospital. We excluded deceased patients for three reasons: (1) adherence to CMS guidelines (including the need for a discharge PI stage that is not available in deceased patients), (2) potential bias of the computational model toward learning characteristics of deceased patients instead of HAPI, and (3) weakness and fragility in patients who have terminal illness result in PI occurrence and do not reflect poor nursing care quality. We conducted an experiment that included deceased patients and observed that some deceased patients were classified as HAPI when they were not HAPI cases.

We found that some HAPI cases lacked PI staging events and yet contained PI keywords in their notes. Thus, the *EHAPI* also checked the PI-related keywords that occurred in notes later than 24 hours after admission. We scanned all stays for PI-related keywords mentioned in notes later than 24 hours after admission; if present, we considered these cases to be HAPI cases. We used negation detection and analyzed the notes of these cases to ensure that the keywords were not spurious (Multimedia Appendix 1). Other stays constituted the control

group. Figure 2 provides the flowchart for the *EHAPI* case definition process.

To ensure the generalizability of the *EHAPI* definition, Multimedia Appendix 1 shows the HAPI-related SNOMED and ICD-10 codes used in many clinical information systems. However, the keywords for notes may need to be tailored to each hospital system. For details of the PI lists, keywords, and mappings in MIMIC-III used for the *EHAPI*, CANTRIP, Cramer, and Sotoodeh definitions, we refer the reader to Multimedia Appendix 1.

Figure 2. Flowchart for the Emory hospital-acquired pressure injury (*EHAPI*) definition. Common inclusion criteria across existing definitions and *EHAPI* are the presence of notes, patients aged 15 years and older, discharge time after admission time, and no pressure injury (PI) diagnosis on admission. D: dimension; DTI: deep tissue injury; HAPI: hospital-acquired pressure injury; MIMIC III: Medical Information Mart for Intensive Care III.



Assessing Impact of HAPI Labels on Classification Performance

We compared the 3 existing HAPI definitions for HAPI classification in MIMIC-III with our *EHAPI* definition. One systematic review [21] study looked at data-driven models for PI prediction and risk assessment and concluded that many of these predictive models were difficult to compare because they were not externally validated and did not use the same data set.

Event Time Stamp Definition

For each hospital stay, we identified an event time stamp for feature construction. The idea is that for HAPI cases, HAPI-related information is not directly or indirectly present in the features (ie, target or label leakage). Similarly, for non-HAPI cases, we prevented biasing the classifier from predicting longer note durations that would be associated with non-HAPI stays. The event time stamp for HAPI cases is the time of the first PI stage assessment that occurred later than 24 hours after admission. The assessment is the earliest time stamp of either the PI staging chart event or the mention of any of the defined PI keywords in the notes.

For control stays, we matched the non-HAPI duration distribution with the HAPI duration distribution. We modeled the duration of the notes in case stays (the time difference between the earliest note and event time stamp) as a random variable. We used the *scikit-learn* [22] package to learn the density distribution for this random variable (ie, the estimated distribution with the smallest chi-square score). We then sampled the duration from this estimated distribution for the allowed note length for control stays. Each sampled duration pairs with a true duration length by preserving the ranked order (ie, the fifth smallest duration of sampled length and true length paired with each other). The minimum sampled length and true duration length then serve as the event time stamp (ie, earliest note+sampled length) for the control stay. Thus, if the sampled event time stamp exceeds the stay duration, then the event time stamp is the original stay discharge time.

Notes for HAPI Classification

Hospital stay features are based on patient notes. Machine learning models then only use notes with a time stamp before the event time stamp, or notes of interest, as features. If there were no notes of interest, the stay was excluded from the experiments. The notes of interest were then concatenated into a single document. This minimized the potential for label or target leakage, where HAPI-related information was directly or indirectly present in the features. For instance, defined PI-related keywords do not appear in the concatenated document. Similarly, the feature construction excludes notes after the first staging assessment, thereby preventing implicit PI-related words.

Thus, feature construction excludes all elements discussed in the definition of the HAPI.

Classifiers

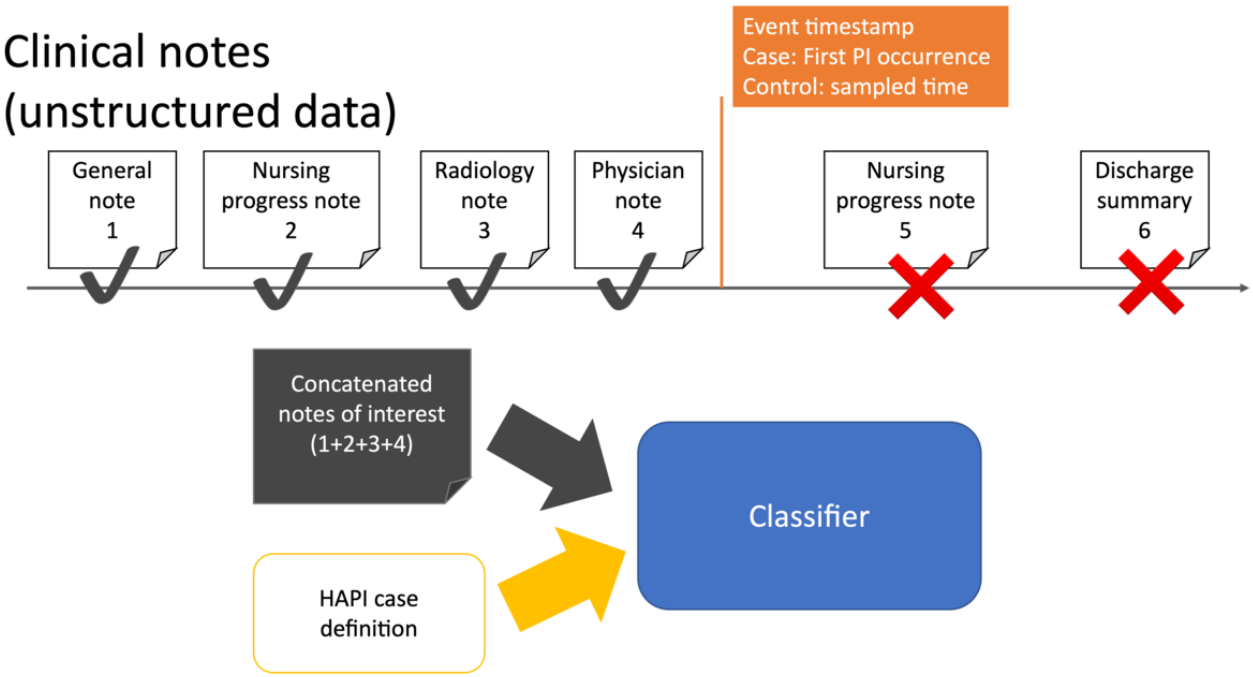
We chose 2 classifiers to demonstrate that the relative impacts of HAPI labels from the case or control definitions are independent of classifier choice. We chose gradient boosting, a tree-based classifier, and a sequential neural network–based classifier. The latter consisted of input word embedding learned from the features of each definition, a global max pooling layer, and several dense layers. The selected classifiers offer superior performance compared with other tested classifiers (ie, decision tree, logistic regression, support vector machine, multilayer perceptron, random forest, or AdaBoost).

The term frequency–inverse document frequency vector of the abridged notes (described in the aforementioned section) is the feature vector of each stay for the tree-based classifier (with a 5000-word vocabulary). The sequential neural network model uses a sequence of 800 words for each document. The 4 different HAPI definitions (ie, CANTRIP, Cramer, Sotoodeh, and *EHAPI*) used the same features to yield unbiased model performance comparisons with different definitions.

Train-Test Compositions and Evaluation Metrics

Because HAPI criteria differed across definitions, the samples for the prior papers were different (eg, *EHAPI* discards deceased patients, but others have it either as a case or control). Nevertheless, there were considerable case overlaps across the samples (Figure 3). For a valid comparison, we created 10 different test sets consisting of three parts: (1) consensus HAPI case stays where the definitions agreed, (2) randomly subsampled consensus HAPI control stays, and (3) manually annotated stays where the definitions disagreed. For the latter stays, our nursing experts, a coauthor (WZ), and her nurse colleague (Deborah Silverstein, DNP) assessed and labeled 97 patient stays for HAPI based on the EHR data. The 97 stays constructed each constituent subset proportional to the total size of the differently labeled stay subset (Table S1 in Multimedia Appendix 1). Annotation relies not only on nursing guidelines but also on nursing experience and case discussion between the 2 nurses. Furthermore, one of our nurse annotators (Deborah Silverstein, DNP) was unaware of the *EHAPI* criteria and labeled the samples from a clinical practitioner's perspective. Out of the 97 admissions, our nursing experts marked 19 as HAPI.

Figure 3. An illustrative example of the feature construction used to train the classifier including determination of the notes of interests using the event time stamp. Feature construction excludes notes 5 and 6 since they are after the event time stamp. HAPI: hospital-acquired pressure injury; PI: pressure injury.



The manually annotated subset was augmented with consensus stays. There were 219 HAPI cases identified by the 4 definitions, which were included in the 10 test samples. The remaining 3620 non-HAPI stays were randomly sampled from the 41,241 admissions where all 4 definitions agreed on the label. Each test set contained 3936 stays and 7% HAPI prevalence. The main difference between each test set was the 3620 randomly sampled non-HAPI consensus stays, as the 97 manually annotated stays and 219 consensus HAPI cases were present in all test sets.

For each definition, the training samples were the remaining eligible stays that were not in the shared test set. As an example, because CANTRIP did not exclude stays with deceased patients, we included these in the training sample. The training labels were set using definition-specific HAPI criteria. Therefore, although the test samples and labels were the same, the classifier trained for each definition had a different training set and definition-specific label. Figure 3 illustrates the overall process for training the classifier, including the feature construction and label determination.

For the experiments, 5-fold cross-validation of the training set determined the best classifier hyperparameters, as shown in Table S2 in Multimedia Appendix 1. The test performance was the average of 10 different test and training data partitions. Given the unbalanced classes, we report both area under the precision-recall curve (AUPRC) and area under the receiver operating characteristic curve (AUROC).

Ethical Considerations

The patients were not explicitly recruited to acquire the data used in this work. The MIMIC-III data set has been deidentified through elimination of attributes revealing patients' identity.

Approval for data collection, processing, and release for the MIMIC-III database was granted by the Institutional Review Boards of the Beth Israel Deaconess Medical Center (Boston, United States) and Massachusetts Institute of Technology (Cambridge, United States).

Results

Consistency of MIMIC-III Data Sources for PI Hospital Stays

Even without consideration of patient attributes or timing, evidence shows conflicts among data sources for identifying PI case hospital stays in MIMIC-III. Figure 1 presents an UpSet plot that summarizes the intersection of PI-related information across the 4 data sources (ie, procedure codes, diagnosis codes, notes, and chart events) for stays with at least one data source indication of PIs (7908 total stays). The data source bar charts (bottom left side) plot the cardinality (or size) of the number of stays with the data source indication. Chart events were the most common indicator (6243/7908, 78.95%), whereas procedure codes only appeared in 3.53% (279/7908) of the stays. The bar charts along the x-axis plot the size of the intersections among the observed set combinations. The results demonstrate limited consensus among the data sources, as only 0.25% (20/7908) of the stays had PI documentation across all data sources. Even agreement among ≥3 data sources was relatively low, with 945 stays (927+9+9). This can be contrasted with 49.34% (3902/7908) and 9.71% (768/7908) of the stays containing only an indication in the chart events and notes, respectively.

Analyzing Differences in HAPI Case Definitions

On the basis of the 4 HAPI case definitions, there are 8 dimensions in which the criteria diverge. The exclusion criteria encompass deceased patients (D1), minimum age (D2), and the amount of time to ascertain PIs on admission (D3). The determination of HAPI includes the minimum PI stage (D4), consideration of deep tissue injury or unstageable events (D5), use of PI-specific keywords in the notes (D6), calculation of deteriorating or new PIs (D7), and use of ICD-9 codes (D8). Table 1 summarizes the decisions along these 8 dimensions for the 4 different definitions. As can be observed from the table, *EHAPI* definition excludes the deceased entirely from case or control and ascertains whether the PI deteriorated or newly developed. Both the Cramer and Sotoodeh definitions yielded substantially lower estimates of HAPI prevalence, whereas CANTRIP had the highest prevalence at 8.46% (4261/50,376).

An UpSet plot capturing the overlap between HAPI stays across the 4 definitions is shown in Figure 4. Only 4.63% (219/4731) of HAPI stays shared 4 definitions. CANTRIP had the highest number of unique positive stays (n=1134), arising from considering PI stages above 1 and deep tissue injury and unstageable events as positives. We observed 315 stays unique to *EHAPI*, attributed to the cutoff period (24 vs 48 with CANTRIP). *EHAPI* had the highest overlap, 53.98% (2554/4731) with CANTRIP, followed by Cramer at 22.53% (1066/4731) and Sotoodeh at 17.84% (844/4731). The details of the number of PIs identified using notes, staging, and ICD-9 codes for each definition are provided in Multimedia Appendix 1.

Table 1. Definition properties and compositions along the 8 criteria dimensions (Ds).

Definition	D1[a]	D2[b]	D3[c]	D4[d]	D5[e]	D6[f]	D7[g]	D8[h]	Cases, n (%)
EHAPI[i] (n=44,823)	Yes	15	24 h	2	Yes	Yes	Yes	No	2976 (6.64)
CANTRIP[j] (n=50,376) [10]	No	15	48 h	1	Yes	Yes	No	No	4261 (8.46)
Cramer (n=50,276) [6]	No	18	24 h	2	No	No	No	No	1572 (3.13)
Sotoodeh (n=50,276) [9]	No	18	N/A[k]	N/A	N/A	Yes	No	Yes	1027 (2.04)

[a]D1 denotes the decision of whether to exclude deceased.

[b]D2 refers to the minimum age in years.

[c]D3 indicates the cutoff period for determining preexisting pressure injury (PI).

[d]D4 characterizes the minimum numerical PI stage.

[e]D5 signifies whether deep tissue injury or unstageable PI staging chart events are hospital-acquired pressure injury (HAPI).

[f]D6 represents whether PI keywords present in notes are considered an HAPI event.

[g]D7 designates whether the criteria captured worsening or newly developed PI.

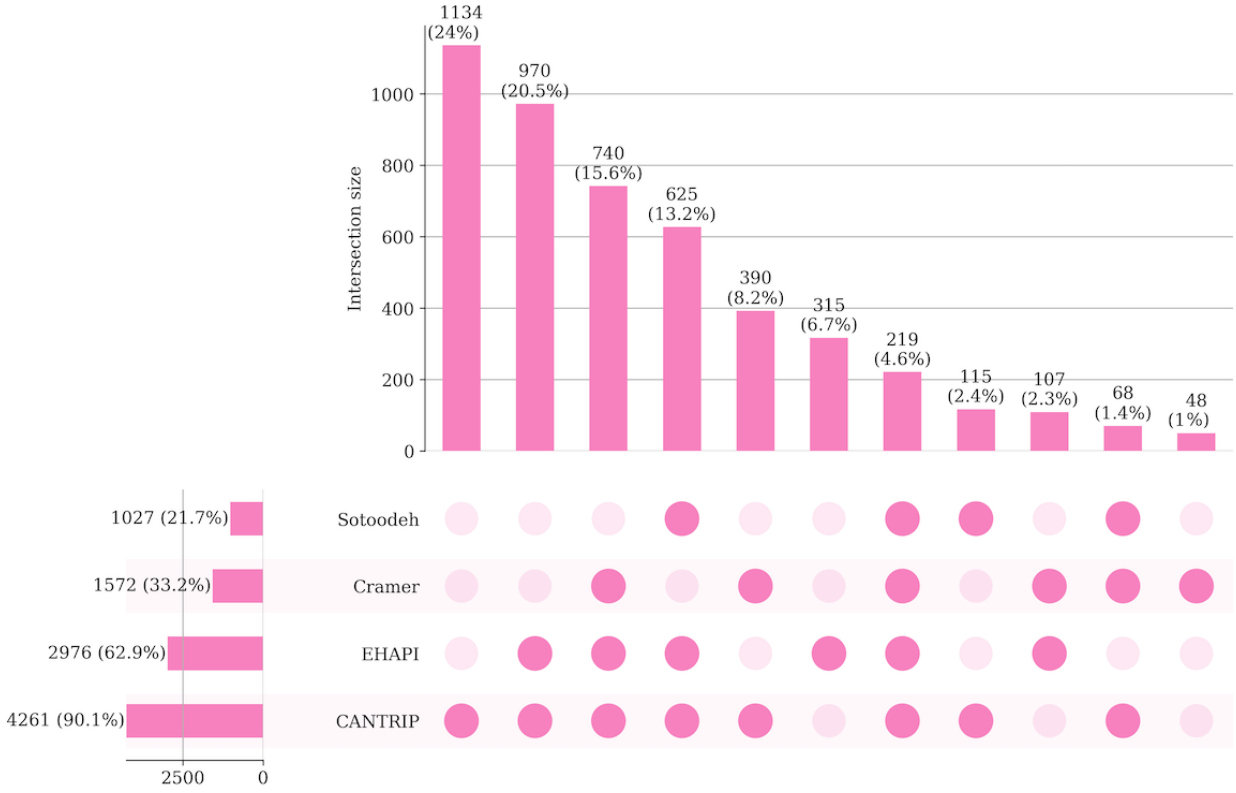
[h]D8 captures whether International Classification of Diseases 9 codes use HAPI for identification.

[i]*EHAPI*: Emory hospital-acquired pressure injury.

[j]CANTRIP: recurrent additive network for temporal risk prediction.

[k]N/A: not applicable.

Figure 4. Overlap of hospital-acquired pressure injury (HAPI) stays across the 4 definitions. CANTRIP: recurrent additive network for temporal risk prediction; *EHAPI*: Emory hospital-acquired pressure injury.



Impact of HAPI Labels on Classification Performance

We evaluated the performance of the gradient boosting and sequential neural network classifiers trained on labels determined by the 4 HAPI definitions by using AUPRC and AUROC for each case. Table 2 presents the results for the 10 described test sets. Table S3 in Multimedia Appendix 1 summarizes the performance based on the test label source (ie, nurse or consensus). Classifiers trained on the *EHAPI* criteria performed better than those trained on other 3 criteria with an improvement in AUROC up to 0.03 and in AUPRC up to 0.11. A 1-sided paired *t* test (1-tailed) between *EHAPI* and the next best performing definition (CANTRIP) resulted in a *P* value of <.001 for AUPRC and AUROC for the better-performing gradient boosting classifier and machine epsilon for other classifiers and definitions (except neural networks and CANTRIP), demonstrating the merits of the *EHAPI* definition. Further analysis of the models' performance stability and the most important words in each setting are provided in Figure S6 and Table S4 in Multimedia Appendix 1. A GitHub repository [23] contains scripts for these experiments, the generation of the stay labels, and the other presented results.

Table 2. Classifiers' performance for the 4 hospital-acquired pressure injury definitions in Medical Information Mart for Intensive Care III over 10 test sets. The results represent the average across test sets with the SD in parenthesis and the *P* value.

Definition	Gradient boosting				Neural networks			
	AUPRC[a]		AUROC[b]		AUPRC		AUROC	
	Mean (SD)	*P* value	Mean (SD)	*P* value	Mean (SD)	*P* value	Mean (SD)	*P* value
EHAPI[c]	0.46 (0.015)	N/A[d]	0.90 (0.003)	N/A	0.44 (0.015)	N/A	0.88 (0.004)	N/A
CANTRIP[e] [10]	0.44 (0.017)	≤.001	0.89 (0.003)	≤.001	0.41 (0.010)	≤.001	0.88 (0.005)	.02
Cramer [6]	0.35 (0.015)	≤.001	0.87 (0.005)	≤.001	0.38 (0.022)	≤.001	0.86 (0.006)	≤.001
Sotoodeh [9]	0.33 (0.015)	≤.001	0.87 (0.004)	≤.001	0.35 (0.015)	≤.001	0.86 (0.006)	≤.001

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

[b]AUROC: area under the receiver operating characteristic curve.

[c]*EHAPI*: Emory hospital-acquired pressure injury.

[d]N/A: not applicable.

[e]CANTRIP: recurrent additive network for temporal risk prediction.

Discussion

Principal Findings

Given the low concurrence of PI between data sources, any HAPI classification requires careful reconciliation of conflicts between data sources. On the basis of discussions with our nursing collaborators (WZ, RN, PhD; Deborah Silverstein, RN, DNP; and RLS, RN, DNP), we prioritized data source reliability as (1) chart events, (2) notes, and (3) diagnosis codes. Charting events were least likely to have a false positive and had better coverage than the other 2 data sources. The nurses indicated that PI indications from notes have false positives, as keywords are preceded by a negative word (ie, no PI), or denote suggestions for PI prevention. Diagnosis codes include only the most prominent diagnoses and might include diagnoses of earlier admissions. In addition, their lack of time stamps prevents investigation of the deterioration condition of the HAPI. Because procedure codes are not specific and inconsistent with other PI sources, we excluded them from the *EHAPI* definition.

As shown in Table 2, the classifiers trained using the *EHAPI* definition achieved the best performance. Moreover, the AUROC of the resulting classifiers from 4 definitions were consistently high (≥0.86). The high AUROC is consistent with the CANTRIP results (AUROC of 0.87) [10] and Sotoodeh results (AUROC of 0.95) [9]. However, the AUPRC remains unacceptable, with the highest performance achieved by gradient boosting (0.46). These values are consistent with the existing literature, as CANTRIP reported precision and recall of 0.42 and 0.71, respectively [10], and Cramer reported precision and recall of 0.09 and 0.71, respectively [6]. This illustrates that to identify the HAPI cases, the computational model generates a sizeable portion of false positives.

Limitations and Future Work

CMS-defined guidelines specify that HAPI are only newly developed, unhealed, or deteriorated PIs. Unfortunately, this involves matching admission and discharge PIs, as a patient may be admitted with >1 PI and discharged with more or fewer PIs. The deterioration condition describes each PI individually. However, given the limited data in the event table of MIMIC-III, our case criteria assume that stays are associated with only one PI. Further analysis of multiple possible PI locations yielded better grouping. However, unless skin assessments at admission and discharge are documented in a structured format, matching PIs is difficult. Ideally, the “deteriorated PI” criterion applies to positive PI samples using patient notes as well. However, information on the PI stage is difficult to obtain from notes and, thus, is not implemented in the current case definition. We plan to study the HAPI in other data sets that have better PI documentation practices to fully understand the impact of multiple PIs.

Another limitation of our study is the use of a simple negation detection algorithm to identify false positives occurring with positive PI mentions in the clinical notes. The keyword list disregarded structure matches such as “bedsore: none,” and the negation detection mainly captures instances of text that mentioned “no bedsore observed.” However, instances of negation in more complex textual descriptions may be missed, thus creating false positives in the identified 2976 HAPI stays. A manual inspection of the 1175 case stays labeled through the PI keyword mentions route is left for future work.

Enhancing the manually labeled samples in the 10 test sets beyond the 97 randomly selected ones is another avenue for future research. The small curated set was not large enough for stand-alone analysis, as it yielded large performance variations across the test sets. Unfortunately, it was labor intensive for our nursing annotators to annotate the samples; thus, further annotation is beyond the scope of this work.

In addition, we note that our assessment of the impact of the HAPI definition is based only on MIMIC-III. Furthermore, MIMIC-III contains data collected only in critical care settings. To better understand the performance implications of the HAPI definition, applying implications to other settings, such as the general care units, as well as other health care systems, is needed. We plan to apply these *EHAPI* criteria to define HAPI in more data sets.

In addition to the focus on critical care stays, the MIMIC-III has unique demographic characteristics, such as predominantly Caucasian. We plan to test the generalizability and impact of the *EHAPI* case definition against more data sets with diverse demographics including higher percentages of African American, Asian, and Hispanic individuals or different insurance compositions.

A recent systematic review on the utility of decision support systems for PI management concluded that their adoption in practice has clinical significance in terms of reducing PI incidence and prevalence, but statistical significance was not observed [24]. This emphasizes the importance of studying practical challenges in the adoption of data-driven PI methods by nurses. Moreover, the practical deployment of a computational model necessitates a higher AUPRC to prevent false alarms. Thus, an open question is whether the integration of other patient information in addition to clinical notes, such as physiological measurements, patient demographics, and medications, yields better predictive performance.

Conclusions

An accurate definition of HAPI based on clinical data is critical for automating nursing quality metrics and for valid comparisons of HAPI machine learning models. However, one of the major challenges is the inconsistency of the PI indicators across various data sources. We demonstrate the lack of congruency between the 3 existing HAPI definitions for MIMIC-III and highlight the gaps between each definition and the CMS and AHRQ regulatory guidelines. We then created a refined definition, the *EHAPI*, that more closely reflects the regulatory guidelines. Our experimental results using 2 different classifiers illustrate the impact of the definition on the predictive performance when evaluated on an unseen combination of a small, manually labeled set by 2 nurse annotators and a random sample of the consensus set (ie, all 4 definitions agree on the labels). This reinforces the need for a high-quality standardized HAPI definition, as the *EHAPI* achieves a better predictive performance across multiple test sets.

Acknowledgments

The authors express their deepest gratitude to the following nurse collaborators: Cynthia A Oster, PhD, RN, APRN, MBA, ACNS-BC, ANP, FAAN, Nurse Scientist for Patient Safety—Emory Healthcare, Adjunct Assistant Professor—Nell Hodgson Woodruff School of Nursing, Emory University; and Deborah Silverstein, DNP, APRN, FNP-C, Instructor, Nell Hodgson Woodruff School of Nursing, Emory University. The authors also thank the anonymous reviewers for their suggestions and comments.

This research was supported by the National Library of Medicine of the National Institutes of Health under the award number R01LM013323-01.

Data Availability

Medical Information Mart for Intensive Care III (MIMIC-III) data can be downloaded from the PhysioNet webpage after completing the required Collaborative Institutional Training Initiative, data or human subjects research training. The scripts used to preprocess the data and obtain the presented results can be found in the GitHub repository.

Authors' Contributions

All authors conceptualized the study. MS designed and conducted the data analysis and generated the results. MS conducted the data preprocessing. MS is responsible for the integrity of the work. MS drafted the paper. WZ and RLS provided nursing expertise throughout the study and provided recommendations for study parameters. All authors participated in the writing and revising of the manuscript. All aspects of the study (design; management, analysis, and interpretation of data; writing of the report; and decision to publish) were led by the authors. All authors have read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

The supplementary material containing the Medical Information Mart for Intensive Care III details for the Emory hospital-acquired pressure injury definition, additional comparisons between the 4 case definitions, and further experimental results related to this study.

[DOCX File , 1895 KB - medinform_v11i1e40672_app1.docx]

References

1. Romanelli M, Clark M, Cherry G, Colin D, Defloor T. Science and Practice of Pressure Ulcer Management. Cham: Springer; 2006.
2. Prevention and Treatment of Pressure Ulcers Clinical Practice Guideline. Cambridge, United Kingdom: Cambridge Media; 2014.
3. NPIAP PrU Awareness Fact Sheet. NPIAP. URL: https://cdn.ymaws.com/npiap.com/resource/resmgr/npiap_pru_awareness_fact_she.pdf [accessed 2021-04-20]
4. Levinson D, General I. Adverse events in hospitals: national incidence among Medicare beneficiaries. Department of Health and Human Services Office of the Inspector General. 2010. URL: https://oig.hhs.gov/oei/reports/oei-06-09-00090.pdf [accessed 2022-12-22]
5. Ben-Assuli O, Sagi D, Leshno M, Ironi A, Ziv A. Improving diagnostic accuracy using EHR in emergency departments: a simulation-based study. J Biomed Inform 2015 Jun;55:31-40 [FREE Full text] [doi: 10.1016/j.jbi.2015.03.004] [Medline: 25817921]
6. Cramer EM, Seneviratne MG, Sharifi H, Ozturk A, Hernandez-Boussard T. Predicting the incidence of pressure ulcers in the intensive care unit using machine learning. EGEMS (Wash DC) 2019 Sep 05;7(1):49 [FREE Full text] [doi: 10.5334/egems.307] [Medline: 31534981]
7. Cox J, Schallom M, Jung C. Identifying risk factors for pressure injury in adult critical care patients. Am J Crit Care 2020 May 01;29(3):204-213. [doi: 10.4037/ajcc2020243] [Medline: 32355967]
8. Ranzani OT, Simpson ES, Japiassú AM, Noritomi DT, Amil Critical Care Group. The challenge of predicting pressure ulcers in critically ill patients. A multicenter cohort study. Ann Am Thorac Soc 2016 Oct;13(10):1775-1783. [doi: 10.1513/AnnalsATS.201603-154OC] [Medline: 27463839]
9. Sotoodeh M, Gero ZH, Zhang W, Hertzberg VS, Ho JC. Pressure ulcer injury in unstructured clinical notes: detection and interpretation. AMIA Annu Symp Proc 2020;2020:1160-1169 [FREE Full text] [Medline: 33936492]

10. Goodwin TR, Demner-Fushman D. A customizable deep learning model for nosocomial risk prediction from critical care notes with indirect supervision. J Am Med Inform Assoc 2020 Apr 01;27(4):567-576 [FREE Full text] [doi: 10.1093/jamia/ocaa004] [Medline: 32065628]
11. Johnson AE, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016 May 24;3:160035 [FREE Full text] [doi: 10.1038/sdata.2016.35] [Medline: 27219127]
12. Lundberg S, Lee SI. A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017 Presented at: NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems; Dec 4 - 9, 2017; Long Beach California USA.
13. Long-term care hospital quality reporting program measure calculations and reporting user's manual version 3.1. Centers for Medicare & Medicaid Services. 2019 Oct. URL: https://tinyurl.com/4vkscbvj [accessed 2021-05-01]
14. CMS guideline for LTCH Quality reporting. CMS. URL: https://tinyurl.com/26ct2yfs [accessed 2021-04-07]
15. Section M: Skin Conditions (Pressure Ulcer/Injury). Centers for Medicare and Medicaid Services. 2018 Sep 4. URL: https://tinyurl.com/6xhc83fr [accessed 2021-04-05]
16. Preventing pressure ulcers in hospitals 5. How do we measure our pressure ulcer rates and practices? Agency for Healthcare Research and Quality. URL: https://www.ahrq.gov/patient-safety/settings/hospital/resource/pressureulcer/tool/put5.html [accessed 2021-04-05]
17. Guidelines. The National Pressure Injury Advisory Panel. URL: https://npiap.com/page/Guidelines [accessed 2021-05-01]
18. Song W, Kang M, Zhang L, Jung W, Song J, Bates DW, et al. Predicting pressure injury using nursing assessment phenotypes and machine learning methods. J Am Med Inform Assoc 2021 Mar 18;28(4):759-765 [FREE Full text] [doi: 10.1093/jamia/ocaa336] [Medline: 33517452]
19. Hyun S, Moffatt-Bruce S, Cooper C, Hixon B, Kaewprag P. Prediction model for hospital-acquired pressure ulcer development: retrospective cohort study. JMIR Med Inform 2019 Jul 18;7(3):e13785 [FREE Full text] [doi: 10.2196/13785] [Medline: 31322127]
20. NPIAP Pressure Injury Stages. The National Pressure Injury Advisory Panel. URL: https://npiap.com/page/PressureInjuryStages [accessed 2021-03-27]
21. Jiang M, Ma Y, Guo S, Jin L, Lv L, Han L, et al. Using machine learning technologies in pressure injury management: systematic review. JMIR Med Inform 2021 Mar 10;9(3):e25704 [FREE Full text] [doi: 10.2196/25704] [Medline: 33688846]
22. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in python. J Mach Learn Res 2011;12:2825-2830 [FREE Full text]
23. GitHub. URL: https://github.com/manisci/EHAPI [accessed 2023-02-07]
24. Araujo SM, Sousa P, Dutra I. Clinical decision support systems for pressure ulcer management: systematic review. JMIR Med Inform 2020 Oct 16;8(10):e21621 [FREE Full text] [doi: 10.2196/21621] [Medline: 33064099]

Abbreviations

AHRQ: Agency for Healthcare Research and Quality
AUPRC: area under the precision-recall curve
AUROC: area under the receiver operating characteristic curve
CANTRIP: recurrent additive network for temporal risk prediction
CMS: Center for Medicare and Medicaid
DOE: date of event
EHAPI: Emory hospital-acquired pressure injury
EHR: electronic health record
HAPI: hospital-acquired pressure injury
ICD: International Classification of Diseases
MIMIC-III: Medical Information Mart for Intensive Care III
NPIAP: The National Pressure Injury Advisory Panel
PI: pressure injury

Edited by C Lovis; submitted 30.06.22; peer-reviewed by S Hyun, X Luo, J Li; comments to author 25.09.22; revised version received 24.12.22; accepted 14.01.23; published 23.02.23.

Please cite as:
Sotoodeh M, Zhang W, Simpson RL, Hertzberg VS, Ho JC
A Comprehensive and Improved Definition for Hospital-Acquired Pressure Injury Classification Based on Electronic Health Records: Comparative Study
JMIR Med Inform 2023;11:e40672
URL: https://medinform.jmir.org/2023/1/e40672
doi:10.2196/40672
PMID:36649481

Original Paper

Identification of Postpartum Depression in Electronic Health Records: Validation in a Large Integrated Health Care System

Jeff Slezak[1], MS; David Sacks[1,2], MD; Vicki Chiu[1], MS; Chantal Avila[1], MA; Nehaa Khadka[1], MPH; Jiu-Chiuan Chen[2], MD, SCD; Jun Wu[3*], PhD; Darios Getahun[1,4*], MD, PhD

[1]Kaiser Permanente Southern California, Pasadena, CA, United States

[2]Keck School of Medicine, University of Southern California, Los Angeles, CA, United States

[3]Program in Public Health, Susan and Henry Samueli College of Health Sciences, University of California, Irvine, CA, United States

[4]Kaiser Permanente Bernard J. Tyson School of Medicine, Pasadena, CA, United States

[*]these authors contributed equally

Corresponding Author:
Jeff Slezak, MS
Kaiser Permanente Southern California
100 S. Los Robles Ave
Pasadena, CA, 91101
United States
Phone: 1 626 564 3477
Email: Jeff.M.Slezak@kp.org

Abstract

Background: The accuracy of electronic health records (EHRs) for identifying postpartum depression (PPD) is not well studied.

Objective: This study aims to evaluate the accuracy of PPD reporting in EHRs and compare the quality of PPD data collected before and after the implementation of the *International Classification of Diseases, Tenth Revision* (*ICD-10*) coding in the health care system.

Methods: Information on PPD was extracted from a random sample of 400 eligible Kaiser Permanente Southern California patients' EHRs. Clinical diagnosis codes and pharmacy records were abstracted for two time periods: January 1, 2012, through December 31, 2014 (*International Classification of Diseases, Ninth Revision* [*ICD-9*] period), and January 1, 2017, through December 31, 2019 (*ICD-10* period). Manual chart reviews of clinical records for PPD were considered the gold standard and were compared with corresponding electronically coded diagnosis and pharmacy records using sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). Kappa statistic was calculated to measure agreement.

Results: Overall agreement between the identification of depression using combined diagnosis codes and pharmacy records with that of medical record review was strong (κ=0.85, sensitivity 98.3%, specificity 83.3%, PPV 93.7%, NPV 95.0%). Using only diagnosis codes resulted in much lower sensitivity (65.4%) and NPV (50.5%) but good specificity (88.6%) and PPV (93.5%). Separately, examining agreement between chart review and electronic coding among diagnosis codes and pharmacy records showed sensitivity, specificity, and NPV higher with prescription use records than with clinical diagnosis coding for PPD, 96.5% versus 72.0%, 96.5% versus 65.0%, and 96.5% versus 65.0%, respectively. There was no notable difference in agreement between *ICD-9* (overall κ=0.86) and *ICD-10* (overall κ=0.83) coding periods.

Conclusions: PPD is not reliably captured in the clinical diagnosis coding of EHRs. The accuracy of PPD identification can be improved by supplementing clinical diagnosis with pharmacy use records. The completeness of PPD data remained unchanged after the implementation of the *ICD-10* diagnosis coding.

(JMIR Med Inform 2023;11:e43005) doi:10.2196/43005

KEYWORDS

validation; postpartum depression; electronic health records; pregnancy; health care system; diagnosis codes; pharmacy records; health data; data collection; implementation; eHealth record; depression; mental well-being; women's health

Introduction

Postpartum depression (PPD), major or minor depressive episodes occurring within 12 months after childbirth, is a common obstetric complication in the United States, with a prevalence of 13.2% in 2018 [1]. The American College of Obstetricians and Gynecologists recommends all obstetrics care providers conduct comprehensive screening for PPD and anxiety disorders using a validated instrument for each patient separately during their postpartum visit [2]. Meanwhile, the American Academy of Pediatrics recommended routine PPD screening to be integrated at well-child visits (1-, 2-, 4-, and 6-month infant visits) [3]. The US Preventive Services Task Force also supports the provision of depression screening during postpartum visits, citing moderate net benefits for identifying those affected and recommending referrals to counseling interventions [4]. It is important to identify those with PPD because undetected or untreated depressive episodes can negatively impact the patient and their infant's health and well-being. For instance, about 9% of pregnancy-related deaths were due to mental health conditions [5]. Early PPD was also associated with increased behavior disturbances in the infant [6]. Moreover, other potential risk factors, including a prior history of depression, depression and anxiety episodes during pregnancy, preterm birth and lower infant birth weight, traumatic birth experience, stressful life events during early postpartum, and low social support, have been linked with PPD [7-9].

Health systems previously used the *International Classification of Diseases, Ninth Revision* (*ICD-9*), an official coding system to identify hospital-related diagnoses and procedures in the United States [10]. However, the Kaiser Permanente health systems shifted to using the *International Classification of Diseases, Tenth Revision* (*ICD-10*) codes after October 1, 2015, which has significant improvements over *ICD-9* for many clinical codes [11]. However, Stewart et al [12] concluded that there is a need to perform a validation of diagnosis codes for each mental health condition following the *ICD-10* transition. Colvin et al [13] used a data linkage of national pharmacy records and hospital admission information to identify patients with major depressive episodes in pregnancy but found the use of either source alone to be inadequate.

While there are multiple validated scales to screen for PPD, like the Patient Health Questionnaire (9-item) and the Edinburgh Postnatal Depression Scale, validation of these measures has been performed using *ICD-9* or *ICD-10* diagnostic codes as the gold standard [14,15]. Several studies have also developed machine learning algorithms using electronic health record (EHR) data to create risk-based models and examined whether they can predict PPD in large health care systems, relying on PPD ascertained using *ICD-9* or *ICD-10* codes [16,17]. However, the accuracy of *ICD-9* and *ICD-10* codes as the gold standard in ascertaining PPD has not been established previously. Prior validation of *ICD-9* and *ICD-10* found high positive predictive values (PPVs) for ascertaining general depression (89.7% and 89.5%, respectively), but these were not specific to the postpartum period [18]. This study aimed to assess the validity of ascertaining PPD diagnosis using the EHR from a large integrated health care delivery system, Kaiser Permanente Southern California (KPSC).

Methods

Cohort Selection

We identified a random sample of 400 women with live birth records in the Air Pollution and Pregnancy Complications in Complex Urban Environments (APPCUE) study [19] between January 1, 2008, and December 31, 2018, within KPSC, a large integrated health system. The APPCUE study was a retrospective cohort study conducted in collaboration between KPSC and the University of California, Irvine with access to KPSC's comprehensive EHRs. The APPCUE study included all singleton births at KPSC facilities. The EHRs contain patient-level data from out- and inpatient clinical care, including *ICD-9, Clinical Modification* or *ICD-10, Clinical Modification* diagnosis and procedure codes, as well as pharmacy and laboratory test records. From 236,759 pregnancies during the study period, we excluded pregnancies resulting in nonlive births (n=8422) and patients who were not members from the start of their pregnancy through a 1-year postpartum period (n=70,836) to have a complete medical history for this validation study. Of the remaining 157,501 pregnancies, we selected a random sample of 400. Simple random sampling was used to select 100 patients from groups based on EHR data: those without any diagnostic or pharmacy use record for PPD, those with only a diagnostic code for PPD, those with only a pharmacy record indicating treatment for PPD, and those with both diagnostic and pharmacy indications. Additionally, each sample was evenly split (50 each) between the *ICD-9* diagnosis code era (date of delivery 2012-2014) and the *ICD-10* era (2017-2019).

Outcomes

EHR outcomes were determined by the presence of PPD diagnosis codes in inpatient or outpatient encounters in the 12 months after delivery, new prescription order, or pharmacy dispense for the treatment of PPD. Diagnosis codes during the *ICD-9* coding period were 300.4, 309.0, and 311 and during the *ICD-10* period were F32.9, F33.0, F33.2, F33.3, F33.41, F33.9, F34.1, F43.21, and F53.0. Medications included were bupropion, Celexa, citalopram, Cymbalta, desvenlafaxine, duloxetine, Effexor, escitalopram, fluoxetine, Lexapro, paroxetine, Paxil, Pristiq, Prozac, sertraline, venlafaxine, Wellbutrin, and Zoloft.

Gold standard PPD outcomes were determined by review of health records by trained research personnel, who documented any diagnosis or finding of PPD in the record, including in free-text encounter notes, as well as any prescription given for the treatment of PPD. These included new prescriptions for the treatment of PPD. PPD diagnosis and medication were documented independently, both for the EHR data and the chart review. A mother was considered to have PPD if she had either a diagnosis or a prescription noted in the EHR within 1 year postpartum.

Quality Assurance

Multiple individuals were trained on reviewing charts, and a double chart review was performed at the beginning of data

collection as a training exercise and near the middle and at the end of data collection to verify data quality and consistency. At each point, eight charts were randomly selected for review by two abstractors. In case of disagreement on the findings, abstractors met with the trainer to determine the correct result.

Statistical Analysis

The patient population was described in terms of demographics, smoking status, prenatal care, and birth weight using percentages. These characteristics were also described for the study population of the APPCUE study [19] and all live births among KPSC members and the state of California during the study timeframe. The chi-square test was used to compare the distribution of characteristics in the study sample to the APPCUE population, all KPSC births, and the California birth cohort.

Manual chart review findings were treated as the true PPD status. The sensitivity, specificity, PPV, and negative predictive value (NPV) of the electronic records to identify true PPD status were calculated and presented as a percentage and 95% exact binomial CI. Agreement between electronic records and manual review was calculated using the kappa statistic, which adjusts for agreement expected due to random chance, and its 95% CI. The area under the receiver operating characteristic curve was calculated. Each measure was calculated overall and within the *ICD-9* and *ICD-10* coding eras separately. There was no missing data for PPD status; those without documented PPD diagnosis or medication were taken to not have PPD. For patient characteristics, a missing category was included when presenting the data.

The primary analysis focused on the ability of EHRs to capture PPD, while secondary analyses examined the agreement of diagnosis and prescription records separately. The sample size was selected so that the expected width of the CIs for sensitivity and PPV would be at most 10% for the full sample and 13% for the *ICD-9* and *ICD-10* periods if the true sensitivity and PPV were 80%. Higher sensitivity and PPV would yield narrower CIs. The STARD (Standards for Reporting Diagnostic Accuracy Studies) guidelines were followed. All analyses were performed in SAS version 9.4 (SAS Institute).

Ethics Approval

The study was approved by the institutional review board of KPSC and received a waiver for informed consent (IRB 12110).

Results

Cohort Selection

Table 1 shows the distribution of the APPCUE study cohort as well as the overall KPSC birth cohort during the study period. Nearly half (194/400, 48.5%) were Latina, most (379/400, 94.8%) received prenatal care starting in the first trimester, and most (354/400, 88.5%) delivered at 37 weeks of gestation or later. The study sample generally has very similar characteristics to the APPCUE study cohort overall and all KPSC births during the period, though there are some differences relative to all deliveries in the state of California, notably a higher percentage of non-Hispanic White mothers (113/400, 28.3% vs 372,037/2,874,396, 12.9%), older mothers (259/400, 64.8% age ≥30 years vs 1,465,998/2,874,396, 50.0%), and generally higher educational attainment (199/400, 49.8% with at least a college degree vs 1,047,594/2,874,396, 36.5%).

Table 1. Characteristics of the study sample and women delivered in all Kaiser Permanente Southern California (KPSC) hospitals and the state of California (2012-2014 and 2017-2019).

Characteristics	Chart review sample[a] (N=400), n (%)	APPCUE[b] study population (N=157,501), n (%)	*P* value	All KPSC births (N=236,759), n (%)	*P* value	All California State births[c] (N=2,874,396), n (%)	*P* value
Maternal age (years)			.42		.01		<.001
<20	10 (2.5)	4665 (3.0)		6804 (3.0)		144,945 (5.0)	
20-29	131 (32.8)	56,679 (36.0)		92,203 (40.4)		1,263,453 (44.0)	
30-34	153 (38.3)	54,810 (34.8)		75,633 (33.1)		843,010 (29.3)	
≥35	106 (26.5)	41,347 (26.3)		53,697 (23.5)		622,988 (21.7)	
Race/ethnicity			.31		.29		<.001
Non-Hispanic White	113 (28.3)	39,219 (24.9)		55,218 (24.2)		372,037 (12.9)	
Non-Hispanic Black	32 (8.0)	10,862 (6.9)		16,207 (7.1)		68,195 (2.4)	
Hispanic	194 (48.5)	78,853 (50.1)		117,162 (51.3)		1,356,354 (47.2)	
Asian/Pacific Islander	47 (11.8)	22,783 (14.5)		31,318 (13.7)		213,499 (7.4)	
Others/unknown	14 (3.5)	5784 (3.7)		8432 (3.7)		864,311 (30.1)	
Educational attainment			.36		.20		<.001
Less than high school	9 (2.2)	4355 (2.8)		6925 (3.0)		435,360 (15.1)	
High school graduate	83 (20.8)	35,411 (22.5)		55,598 (24.4)		694,118 (24.1)	
Some college	99 (24.8)	32,616 (20.7)		50,153 (22.0)		558,288 (19.4)	
Bachelor's/associate's degree	126 (31.5)	54,293 (34.5)		75,849 (33.2)		729,896 (25.4)	
Master's degree/above	73 (18.3)	27,388 (17.4)		34,556 (15.1)		317,698 (11.1)	
Missing	10 (2.5)	3438 (2.2)		5256 (2.3)		139,036 (4.8)	
Household income (US $)			.64		.25		—[d]
<30,000	16 (4.0)	5194 (3.3)		8318 (3.6)		—	
30,000-49,999	90 (22.5)	39,969 (25.4)		61,562 (27.0)		—	
50,000-69,999	124 (31.0)	47,864 (30.4)		69,844 (30.6)		—	
70,000-89,999	82 (20.5)	32,486 (20.6)		45,469 (19.9)		—	
≥90,000	88 (22.0)	31,925 (20.3)		42,782 (18.7)		—	
Prenatal care initiation			.52		<.001		<.001
First trimester	379 (94.8)	147,017 (93.3)		199,866 (87.5)		2,386,232 (83.0)	
No or late care	20 (5.0)	9860 (6.3)		26,966 (11.8)		442,493 (15.4)	
Missing	1 (0.2)	624 (0.4)		1505 (0.7)		45,671 (1.6)	
Smoking during pregnancy	23 (5.8)	6420 (4.1)	.09	10,256 (4.5)	.16	46,977 (1.6)	<.001
Gestational age (weeks)			.13		.08		.16
<34	14 (3.5)	3412 (2.2)		4779 (2.1)		66,099 (2.3)	
34-36	32 (8.0)	9865 (6.3)		13,933 (6.1)		180,352 (6.3)	
≥37	354 (88.5)	144,192 (91.5)		209,553 (91.8)		2,624,620 (91.3)	
Missing	0 (0.0)	32 (0.0)		72 (0.0)		3325 (0.1)	

[a]Sample is based on data from KPSC electronic health records 2012-2014 and 2017-2019.

[b]APPCUE: Air Pollution and Pregnancy Complications in Complex Urban Environments.

[c]Data from the natality information of the Center for Disease Control and Prevention [20].

[d]Data not available.

Outcomes

The overall agreement of EHR-identified PPD (based on either a diagnosis or a prescription) with medical record review was high, with a kappa of 84.7% (95% CI 78.8%-90.6%). The EHR identified 281 of 286 cases (sensitivity 98.3%, 95% CI 96.0%-99.4%) while maintaining high specificity (95.0%, 95% CI 88.7%-98.4%), PPV (93.7%, 95% CI 90.3%-96.1%), and NPV (95.0%, 95% CI 88.7%-98.4%). There was little difference in the overall agreement between the *ICD-9* coding era (κ=86.0%, 95% CI 78.0%-94.0%) and the *ICD-10* era (κ=83.4%, 95% CI 74.8%-92.1%; Table 2).

Table 2. Identification of postpartum depression using diagnostic codes and/or pharmacy records–based data sources before and after implementation of the ICD-10 code in the Kaiser Permanente Southern California system in 2015 (N=400).

	TP[a], n	TN[b], n	FP[c], n	FN[d], n	Sensitivity, % (95% CI)	Specificity, % (95% CI)	PPV[e], % (95% CI)	NPV[f], % (95% CI)	Kappa (95% CI)	AUC[g]
Combined electronic diagnosis codes and pharmacy records										
Overall	281	95	19	5	98.3 (96.0-99.4)	83.3 (75.2-89.7)	93.7 (90.3-96.1)	95.0 (88.7-98.4)	0.85 (0.79-0.91)	0.91
2012-2014	141	48	9	2	98.6 (95.0-99.8)	84.2 (72.1-92.5)	94.0 (88.9-97.2)	96.0 (86.3-99.5)	0.86 (0.78-0.94)	0.91
2017-2019	140	47	10	3	97.9 (94.0-99.6)	82.5 (70.1-91.3)	93.3 (88.1-96.8)	94.0 (83.5-98.7)	0.83 (0.75-0.92)	0.90
***ICD-9*[h]/*ICD-10*[i] diagnosis codes only**										
Overall	187	101	13	99	65.4 (59.6-70.9)	88.6 (81.3-93.8)	93.5 (89.1-96.5)	50.5 (43.4-57.6)	0.44 (0.36-0.52)	0.77
2012-2014	94	51	6	49	65.7 (57.3-73.5)	89.5 (78.5-96.0)	94.0 (87.4-97.8)	51.0 (40.8-61.1)	0.45 (0.34-0.56)	0.78
2017-2019	93	50	7	50	65.0 (56.6-72.8)	87.7 (76.3-94.9)	93.0 (86.1-97.1)	50.0 (39.8-60.2)	0.43 (0.32-0.54)	0.76
Pharmacy records only										
Overall	194	108	6	92	67.8 (62.1-73.2)	94.7 (88.9-98.0)	97.0 (93.6-98.9)	54.0 (46.8-61.1)	0.51 (0.43-0.59)	0.81
2012-2014	97	54	3	46	67.8 (59.5-75.4)	94.7 (85.4-98.9)	97.0 (91.5-99.4)	54.0 (43.7-64.0)	0.51 (0.40-0.62)	0.81
2017-2019	97	54	3	46	67.8 (59.5-75.4)	94.7 (85.4-98.9)	97.0 (91.5-99.4)	54.0 (43.7-64.0)	0.51 (0.40-0.62)	0.81

[a]TP: true positive.

[b]TN: true negative.

[c]FP: false positive.

[d]FN: false negative.

[e]PPV: positive predictive value.

[f]NPV: negative predictive value.

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

[h]*ICD-9: International Classification of Diseases, Ninth Revision.*

[i]*ICD-10: International Classification of Diseases, Tenth Revision.*

Electronic diagnosis records alone were not able to accurately identify PPD, only identifying 187 of 286 cases (sensitivity 65.4%, 95% CI 59.6%-70.9%), with low NPV (50.5%, 95% CI 43.4%-57.6%). PPV (93.5%, 95% CI 89.1%-96.5%) and specificity (88.6%, 95% CI 81.3%-93.8%) were high, however (Table 2). Results were similar when using EHR prescription records alone (sensitivity 67.8%, 95% CI 62.1%-73.2%; specificity 94.7%, 95% CI 88.9%-98.0%; PPV 97.0%, 95% CI 93.6%-98.9%; NPV 54.0%, 95% CI 46.8%-61.1%).

Considering only medication data, the reliability of EHR data for identifying prescriptions for PPD was high, with an overall kappa of 92.5% (95% CI 88.8%-96.2%). Agreement was very high in both the *ICD-9* (κ=92.0%, 95% CI 86.6%-97.4%) and *ICD-10* eras (κ=93.0%, 95% CI 87.9%-98.1%; Table 3). Sensitivity, specificity, PPV, and NPV were all at or above 96% (Table 3).

Agreement for *ICD* diagnostic codes between EHR and manual chart review was much lower overall (κ=55.0%, 95% CI 47.1%-62.9%; Table 3). The PPV was high (90.0%, 95% CI 85.0%-93.8%), with sensitivity lower (72.0%, 95% CI 66.0%-77.5%) and specificity and NPV much lower (both 65.0%, 95% CI 58.0%-71.6%; Table 3). Agreement was similar

between the *ICD-9* (κ=58.0%, 95% CI 47.1%-68.9%) and *ICD-10* (κ=52.0%, 95% CI 40.5%-63.5%) eras (Table 3).

Table 3. Identification of postpartum depression based on individual data sources before and after implementation of the ICD-10 code in the Kaiser Permanente Southern California system in 2015 (N=400).

	TP[a], n	[b]TN, n	FP[c], n	FN[d], n	Sensitivity, % (95% CI)	Specificity, % (95% CI)	PPV[e], % (95% CI)	NPV[f], % (95% CI)	Kappa (95% CI)	AUC[g]
***ICD-9*[h]/*ICD-10*[i] diagnosis codes only**										
Overall	180	130	20	70	72.0 (66.0-77.5)	86.7 (80.2-91.7)	90.0 (85.0-93.8)	65.0 (58.0-71.6)	0.55 (0.47-0.63)	0.79
2012-2014	92	66	8	34	73.0 (64.4-80.5)	89.2 (79.8-95.2)	92.0 (84.8-96.5)	66.0 (55.8-75.2)	0.58 (0.47-0.69)	0.78
2017-2019	88	64	12	36	71.0 (62.1-78.8)	84.2 (74.0-91.6)	88.0 (80.0-93.6)	64.0 (53.8-73.4)	0.52 (0.41-0.64)	0.81
Pharmacy records only										
Overall	192	193	8	7	96.5 (92.9-98.6)	96.0 (92.3-98.3)	96.0 (92.3-98.3)	96.5 (92.9-98.6)	0.93 (0.89-0.96)	0.96
2012-2014	96	96	4	4	96.0 (90.1-98.9)	96.0 (90.1-98.9)	96.0 (90.1-98.9)	96.0 (90.1-98.9)	0.92 (0.87-0.97)	0.97
2017-2019	96	97	4	3	97.0 (91.4-99.4)	96.0 (90.2-98.9)	96.0 (90.1-98.9)	97.0 (91.5-99.4)	0.93 (0.88-0.98)	0.96

[a]TP: true positive.

[b]TN: true negative.

[c]FP: false positive.

[d]FN: false negative.

[e]PPV: positive predictive value.

[f]NPV: negative predictive value.

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

[h]*ICD-9*: *International Classification of Diseases, Ninth Revision*.

[i]*ICD-10*: *International Classification of Diseases, Tenth Revision*.

Quality Assurance

During the training process, 8 charts were independently reviewed by two chart abstractors. Their assessments of medication use for PPD agreed for all 8 records (100%), while the assessment of a diagnostic finding agreed for 7 (88%). After training was complete, another 8 records were independently reviewed. All 8 (100%) agreed in their findings for both medications and diagnoses.

Discussion

Principal Findings

This validation study demonstrated the potential to improve the accuracy of PPD case identification from an EHR when using diagnosis codes in conjunction with pharmacy records. The combination of clinical codes and prescription pharmacy records yielded much greater sensitivity and NPV, with no notable loss in specificity or PPV, compared with using either the diagnosis codes or pharmacy records alone. Using either record alone would result in significant undercounting, each missing about one-third of those with PPD, compared to the 95% identified using both together. Furthermore, we observed no significant difference in the *ICD-9* and *ICD-10* codes in terms of ascertaining PPD cases.

We found that electronic records of PPD diagnosis were not a reliable indicator of PPD diagnostic findings identified through chart review, relative to pharmacy records. Pharmacy records have both a sensitivity and specificity much higher than that seen for diagnosis codes.

The quality of data extracted from EHRs for pharmacoepidemiologic research has been proven to be valuable. Although using clinical diagnosis codes for perinatal epidemiology studies has limitations, the use of KPSC's comprehensive pharmacy use records enhances the identification of PPD cases (sensitivity 98.3%, specificity 95.0%, PPV 93.7%, and NPV 95.0%).

While switching from *ICD-9* to *ICD-10* coding created some complexity, we did not see a significant difference in the accuracy of the electronic diagnosis records between the *ICD-9* and *ICD-10* coding eras. This is reassuring, as studies would not need to be limited to one era or the other for the sake of accuracy. Additionally, the prevalence of PPD identified in both periods is essentially the same, suggesting that both *ICD-9* and *ICD-10* coding systems identify patients with PPD at the same

rate, negating any need to adjust prevalence estimates to account for the difference.

Accurate characterization of those with PPD is crucial to performing valid research on this condition. Many researchers rely on electronic records due to a lack of access to detailed patient histories or a lack of time to review these records. Our study suggests that researchers can accurately identify PPD from EHRs using both diagnosis and pharmacy records.

Comparison to Prior Work

Prior research validating diagnosis codes for identifying general depression found the PPV to be similar to that seen in our study (89.7% for *ICD-9* and 89.5% for *ICD-10*), but these were not specific to the postpartum period [18]. These findings highlight the continuing debate regarding the use of diagnosis codes alone for epidemiological studies. Our study concurs with prior findings that the sensitivity and specificity of case ascertainment can be improved by concurrently using both diagnosis and pharmacy records [13]. Therefore, researchers should not rely exclusively on either diagnostic codes or pharmacy records for PPD case ascertainment.

Strengths and Limitations

There are some potential limitations to this study. First, while the KPSC EHR is comprehensive, it may not capture care received outside the system if it is not submitted for reimbursement. Specifically, members may receive mental health counseling from non-KPSC providers, and a PPD diagnosis made in that setting may not be entered into the KPSC medical record, resulting in a potentially missed PPD diagnosis and an underestimate of the sensitivity of diagnosis coding. However, these diagnoses may still be identified during regular clinical care within KPSC, hence limiting the number of potentially missed diagnoses.

Second, misclassification is also possible as variables were ascertained from clinical diagnosis codes and pharmacy record notes. In addition, there is the potential for misclassification of PPD within the data sources if women are unaware of the condition, do not seek medical care, or the diagnosis or treatment is not recorded in the clinical notes. Any completely undocumented cases would result in an underestimate of PPD in the population, though its potential effect on our validation is unknown. Finally, due to the small number of records reviewed in some groups, we were not able to look for differences in medical record accuracy within subsets of the population, including by age and race/ethnicity. If differences are present, this will limit the generalizability of these findings to other populations with different demographics.

Strengths of this study include the comprehensive medical record and chart review conducted to identify PPD in this patient population. The training and validation of the chart review process helped to ensure that the gold standard PPD identification was accurate.

Conclusions

This validation study of PPD that was carried out in a large integrated health care system in Southern California has demonstrated that PPD data ascertainment based on a combination of diagnosis codes and prescription medication records from the EHR is highly accurate for pharmacoepidemiologic studies. Neither diagnosis codes alone nor prescription records alone are sufficient to capture PPD cases.

Acknowledgments

Funding for this research was provided by National Institute of Health grant R01 ES030353-01 to DG (Kaiser Permanente Southern California) and JW (University of California, Irvine). The opinions expressed are solely the responsibility of the authors and do not necessarily reflect the official views of the funding agency. The Air Pollution and Pregnancy Complications in Complex Urban Environments (APPCUE) study team would like to thank Kaiser Permanente members who contributed electronic health information to this study.

Data Availability

Most of the data that support the findings of this study are available on request from the corresponding author. The complete data set is not publicly available due to privacy, institutional approval, and/or ethical restrictions. Study data come from patient electronic health records and birth certificates from the state of California. Data from patient health records cannot be shared without signed confidentiality agreements. Some of the data that support the findings of this study are available from the state of California. Restrictions apply to the availability of these data, which were used under license and approval for this study. Data can be made available by the authors provided that all required approvals are obtained from the departments in the state that oversees the use of state vital records data. To obtain California birth certificate data, researchers can email cphs@chhs.ca.gov or visit their website [21]. Requests for data may be sent to JS (Jeff.M.Slezak@kp.org) and DG (Darios.T.Getahun@kp.org).

Conflicts of Interest

The Air Pollution and Pregnancy Complications in Complex Urban Environments (APPCUE) study team led the design of the study and interpretation of the results. NK, CA, and JCC have no competing interests. JS receives research support from the National Institutes of Health (NIH), Pfizer Inc, Dynavax Technologies, and ALK. DG receives research support from NIH, National Institute of Environmental Health Sciences (NIEHS), Department of Health and Human Services, National Institute of Child Health and Human Development, Patient-Centered Outcomes Research Institute, Garfield Memorial Fund, Bayer AG, and

Hologic, Inc. JW receives research support from NIEHS, the California Air Resources Board, and the Health Effects of Air Pollution Foundation. Kaiser Permanente Southern California (KPSC) led the design of the study and interpretation of the results in collaboration with study team members from the University of California, Irvine (UCI) and the University of Southern California (USC). JS conducted the analyses, which were reviewed by study team members from KPSC, UCI, and USC.

References

1. Bauman BL, Ko JY, Cox S, D'Angelo Mph DV, Warner L, Folger S, et al. Vital signs: postpartum depressive symptoms and provider discussions about perinatal depression - United States, 2018. MMWR Morb Mortal Wkly Rep 2020 May 15;69(19):575-581. [doi: 10.15585/mmwr.mm6919a2] [Medline: 32407302]
2. ACOG Committee. ACOG Committee Opinion No. 757: screening for perinatal depression. Obstet Gynecol 2018 Nov;132(5):e208-e212. [doi: 10.1097/AOG.0000000000002927] [Medline: 30629567]
3. Earls MF, Yogman MW, Mattson G, Rafferty J, Committee on Psychosocial Aspects of Child and Family Health. Incorporating recognition and management of perinatal depression into pediatric practice. Pediatrics 2019 Jan;143(1):e20183259. [doi: 10.1542/peds.2018-3259] [Medline: 30559120]
4. Siu AL, US Preventive Services Task Force (USPSTF), Bibbins-Domingo K, Grossman DC, Baumann LC, Davidson KW, et al. Screening for depression in adults: US Preventive Services Task Force recommendation statement. JAMA 2016 Jan 26;315(4):380-387. [doi: 10.1001/jama.2015.18392] [Medline: 26813211]
5. Pregnancy-related deaths: data from 14 U.S. Maternal Mortality Review Committees, 2008-2017. Centers for Disease Control and Prevention. 2019 Sep 04. URL: https://www.cdc.gov/reproductivehealth/maternal-mortality/erase-mm/mmr-data-brief.html [accessed 2022-02-17]
6. Wrate RM, Rooney AC, Thomas PF, Cox JL. Postnatal depression and child development. A three-year follow-up study. Br J Psychiatry 1985 Jun;146:622-627. [doi: 10.1192/bjp.146.6.622] [Medline: 4016475]
7. Lancaster CA, Gold KJ, Flynn HA, Yoo H, Marcus SM, Davis MM. Risk factors for depressive symptoms during pregnancy: a systematic review. Am J Obstet Gynecol 2010 Jan;202(1):5-14 [FREE Full text] [doi: 10.1016/j.ajog.2009.09.007] [Medline: 20096252]
8. Robertson E, Grace S, Wallington T, Stewart DE. Antenatal risk factors for postpartum depression: a synthesis of recent literature. Gen Hosp Psychiatry 2004;26(4):289-295. [doi: 10.1016/j.genhosppsych.2004.02.006] [Medline: 15234824]
9. Cook N, Ayers S, Horsch A. Maternal posttraumatic stress disorder during the perinatal period and child outcomes: a systematic review. J Affect Disord 2018 Jan 01;225:18-31 [FREE Full text] [doi: 10.1016/j.jad.2017.07.045] [Medline: 28777972]
10. International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM). Centers for Disease Control and Prevention. 2021 Nov 03. URL: https://www.cdc.gov/nchs/icd/icd9cm.htm [accessed 2022-02-17]
11. ICD-10-CM Browser Tool. Centers for Disease Control and Prevention. 2021 Jan 26. URL: https://www.cdc.gov/nchs/icd/icd10cm_browsertool.htm [accessed 2022-02-17]
12. Stewart CC, Lu CY, Yoon TK, Coleman KJ, Crawford PM, Lakoma MD, et al. Impact of ICD-10-CM transition on mental health diagnoses recording. EGEMS (Wash DC) 2019 Apr 12;7(1):14 [FREE Full text] [doi: 10.5334/egems.281] [Medline: 31065557]
13. Colvin L, Slack-Smith L, Stanley FJ, Bower C. Are women with major depression in pregnancy identifiable in population health data? BMC Pregnancy Childbirth 2013 Mar 12;13:63 [FREE Full text] [doi: 10.1186/1471-2393-13-63] [Medline: 23497210]
14. Pereira AT, Bos SC, Marques M, Maia BR, Soares MJ, Valente J, et al. The postpartum depression screening scale: is it valid to screen for antenatal depression? Arch Womens Ment Health 2011 Jun;14(3):227-238. [doi: 10.1007/s00737-010-0178-y] [Medline: 20645114]
15. Smith-Nielsen J, Matthey S, Lange T, Væver MS. Validation of the Edinburgh Postnatal Depression Scale against both DSM-5 and ICD-10 diagnostic criteria for depression. BMC Psychiatry 2018 Dec 20;18(1):393 [FREE Full text] [doi: 10.1186/s12888-018-1965-7] [Medline: 30572867]
16. Hochman E, Feldman B, Weizman A, Krivoy A, Gur S, Barzilay E, et al. Development and validation of a machine learning-based postpartum depression prediction model: a nationwide cohort study. Depress Anxiety 2021 Apr;38(4):400-411. [doi: 10.1002/da.23123] [Medline: 33615617]
17. Betts KS, Kisely S, Alati R. Predicting postpartum psychiatric admission using a machine learning approach. J Psychiatr Res 2020 Nov;130:35-40. [doi: 10.1016/j.jpsychires.2020.07.002] [Medline: 32771679]
18. Fiest KM, Jette N, Quan H, St Germaine-Smith C, Metcalfe A, Patten SB, et al. Systematic review and assessment of validated case definitions for depression in administrative data. BMC Psychiatry 2014 Oct 17;14:289 [FREE Full text] [doi: 10.1186/s12888-014-0289-5] [Medline: 25322690]
19. Sun Y, Li X, Benmarhnia T, Chen J, Avila C, Sacks DA, et al. Exposure to air pollutant mixture and gestational diabetes mellitus in Southern California: results from electronic health record data of a large pregnancy cohort. Environ Int 2022 Jan;158:106888 [FREE Full text] [doi: 10.1016/j.envint.2021.106888] [Medline: 34563749]
20. Natality information: live births. CDC WONDER. URL: https://wonder.cdc.gov/natality.html [accessed 2022-01-09]

21. Committee for the Protection of Human Subjects. California Health and Human Services. URL: https://www.chhs.ca.gov/cphs/ [accessed 2023-01-27]

Abbreviations

APPCUE: Air Pollution and Pregnancy Complications in Complex Urban Environments
EHR: electronic health record
ICD-9: International Classification of Diseases, Ninth Revision
ICD-10: International Classification of Diseases, Tenth Revision
KPSC: Kaiser Permanente Southern California
NPV: negative predictive value
PPD: postpartum depression
PPV: positive predictive value
STARD: Standards for Reporting Diagnostic Accuracy Studies

Edited by G Eysenbach, T Leung; submitted 27.09.22; peer-reviewed by Y Chu, C Calvo-Lobo; comments to author 13.12.22; revised version received 03.01.23; accepted 15.01.23; published 01.03.23.

Please cite as:
Slezak J, Sacks D, Chiu V, Avila C, Khadka N, Chen JC, Wu J, Getahun D
Identification of Postpartum Depression in Electronic Health Records: Validation in a Large Integrated Health Care System
JMIR Med Inform 2023;11:e43005
URL: https://medinform.jmir.org/2023/1/e43005
doi:10.2196/43005
PMID:36857123

Original Paper

Successes and Barriers of Health Information Exchange Participation Across Hospitals in South Carolina From 2014 to 2020: Longitudinal Observational Study

Zhong Li[1], PhD; Melinda A Merrell[2,3], PhD; Jan M Eberth[3,4], PhD; Dezhi Wu[5], PhD; Peiyin Hung[2,3], PhD

[1]Department of Public Administration, School of Health Policy and Management, Nanjing Medical University, Nanjing, China

[2]Department of Health Services Policy and Management, Arnold School of Public Health, University of South Carolina, Columbia, SC, United States

[3]Rural and Minority Health Research Center, Arnold School of Public Health, University of South Carolina, Columbia, SC, United States

[4]Department of Health Management and Policy, Drexel University, Philadelphia, PA, United States

[5]Department of Integrated Information Technology, College of Engineering and Computing, University of South Carolina, Columbia, SC, United States

Corresponding Author:
Peiyin Hung, PhD
Rural and Minority Health Research Center
Arnold School of Public Health
University of South Carolina
220 Stoneridge Dr, Suite 204
Columbia, SC, 29210
United States
Phone: 1 8037779867
Email: hungp@mailbox.sc.edu

Abstract

Background: The 2009 Health Information Technology for Economic and Clinical Health Act sets three stages of Meaningful Use requirements for the electronic health records incentive program. Health information exchange (HIE) technologies are critical in the meaningful use of electronic health records to support patient care coordination. However, HIE use trends and barriers remain unclear across hospitals in South Carolina (SC), a state with the earliest HIE implementation.

Objective: This study aims to explore changes in the proportion of HIE participation and factors associated with HIE participation, and barriers to exchange and interoperability across SC hospitals.

Methods: This study derived data from a longitudinal data set of the 2014-2020 American Hospital Association Information Technology Supplement for 69 SC hospitals. The primary outcome was whether a hospital participated in HIE in a year. A cross-sectional multivariable logistic regression model, clustered at the hospital level and weighted by bed size, was used to identify factors associated with HIE participation. The second outcome was barriers to sending, receiving, or finding patient health information to or from other organizations or hospital systems. The frequency of hospitals reporting each barrier related to exchange and interoperability were then calculated.

Results: Hospitals in SC have been increasingly participating in HIE, improving from 43% (24/56) in 2014 to 82% (54/66) in 2020. After controlling for other hospital factors, teaching hospitals (adjusted odds ratio [AOR] 3.7, 95% CI 1.0-13.3), system-affiliated hospitals (AOR 6.6, 95% CI 3.2-13.7), and rural referral hospitals (AOR 8.0, 95% CI 1.2-53.4) had higher odds to participate in HIE than their counterparts, whereas critical access hospitals (AOR 0.1, 95% CI 0.02-0.6) were less likely to participate in HIE than their counterparts reimbursed by the prospective payment system. Hospitals with greater ratios of Medicare or Medicaid inpatient days to total inpatient days also reported higher odds of HIE participation. Despite the majority of hospitals reporting HIE participation in 2020, barriers to exchange and interoperability remained, including lack of provider contacts (27/40, 68%), difficulty in finding patient health information (27/40, 68%), adapting different vendor platforms (26/40, 65%), difficulty matching or identifying same patients between systems (23/40, 58%), and providers that do not typically exchange patient data (23/40, 58%).

Conclusions: HIE participation has been widely adopted in SC hospitals. Our findings highlight the need to incentivize optimization of HIE and seamless information exchange by facilitating and implementing standardization of health information across various HIE systems and by addressing other technical issues, including providing providers' addresses and training HIE stakeholders to find relevant information. Policies and efforts should include more collaboration with vendors to reduce platform

compatibility issues and more user engagement and technical training and support to facilitate effective, accurate, and efficient exchange of provider contacts and patient health information.

(JMIR Med Inform 2023;11:e40959) doi:10.2196/40959

KEYWORDS

health information exchange; electronic health records; interoperability; meaningful use; hospital

Introduction

Health information exchange (HIE) has great potential to support patient transitions and achieve substantial financial and societal benefits across the fragmented United States health care system [1-3]. According to the 2009 Health Information Technology for Economic and Clinical Health (HITECH) Act, there are three stages of Meaningful Use. The goal of Stage 1 of Meaningful Use from 2009 to 2010 was to establish the US federal government's Meaningful Use incentive programs and to motivate health care professionals and institutions to capture, protect, and electronically store data, thus promoting wide adoptions of electronic health records (EHRs) [4]. Stage 2 of Meaningful Use—which began in 2014—emphasized the use and documentation of advanced clinical processes that guided the information exchange between providers and patients and between providers in the same practice to improve treatment adherence and care coordination. In 2015, the Centers for Medicare & Medicaid Services launched Stage 3 of the Meaningful Use requirements for the EHR Incentive Program. In Stage 3, eligible hospitals must demonstrate the interoperability of EHR systems in different practices with a focus on improving patient outcomes [5], such as improved coordination and efficiency of care by reducing the replication of health care services [6,7]. Previous studies have suggested that HIE implementation can also improve quality of care and holds promise to help achieve the goals of other policies, such as the Hospital Readmission Reduction Program [8,9].

Given the potential benefits, the Centers for Medicare & Medicaid finalized a rule to promote HIE and set exchanging all "necessary health information," including courses of illness, treatment, and discharge goals, with health care providers at the next level of care as a condition of participation in Medicare [10,11]. Although primary care providers strongly agree that meeting the Stage 3 care coordination criteria would improve patients' treatment [12], small, rural, and critical access hospitals (CAHs) were less likely to participate in national networks and state, regional, or local health information organizations (HIOs) than other hospitals as of 2018 [13]. Hospitals having a larger market share or those in less competitive markets had a greater probability of HIE participation [14], and nonprofit and publicly owned hospitals were more likely to participate in HIE than for-profit hospitals [15]. Moreover, larger hospitals were more likely to exchange health information internally than with outside hospitals [16]. However, despite the increasing proportion of hospital engagement, patient health information was not exchanged as needed [17]. Emergency department physicians reported that HIE sometimes disrupted their workload [18]. At the organization level, the rates of HIE upon transfer from psychiatric units lagged that from general medicine or surgery hospitals reported as of 2016 [11]. CAHs have been struggling to keep up with other advanced functions, even when these hospitals had EHRs as of 2018 [19], and physicians in solo practices and nonprimary care specialties also lagged [20]. Besides substantially poorer health care infrastructure [21,22], rural hospitals were least likely to have patient engagement capabilities or clinical information available electronically from outside providers [23]. Difficulties in sending and receiving health information and complex workflows are still the main barriers [12]. Previous research pointed out that developers need to work with health care providers to ensure HIE tools are integrated into existing workflows [24]. Costs of electronic interface development might also be a key barrier to fully integrated HIEs [25].

In South Carolina (SC), state-level efforts to encourage HIE were funded by the US Department of Health and Human Services during this time. Building upon existing statewide data infrastructure, SC became an early adopter of HIE starting in 2008 [26]. However, adverse effects of privacy regulation on the successful implementation and use of HIE also raised wide concerns [3], presenting a crucial obstacle to facilitating the exchange of health information between hospitals and across different health systems or distinct physician practices [27]. Because data on the implementation of HIE and barriers to exchange and interoperability of patient health information across hospitals in SC remains unknown, we pursued this study to provide insight for the state government to enact purposive policies and intensive efforts to promote wider participation in and use of HIE. The anticipated benefits gained from HIE come from the realized exchange and use of patient health information across health care providers; thus, whether these health information technologies are run in a supportive environment raised concerns about the barriers SC hospitals face. Previous studies found that technical and cost issues and privacy and security concerns could hinder HIE implementation during the process of HIE expansion [28,29], even with an increasing adoption of EHRs in the United States [30]. As qualification to participate in HIE does not necessarily lead to the use of HIEs [25], identifying and assessing related obstacles may inform policy efforts to address health care providers' concerns about their HIE use for improved health outcomes when most hospitals have met the criteria of Meaningful Use Stage 3 [27]. As an early adopter of HIE [26], SC, which contains some very rural areas, may have faced unique challenges that might be insightful for the adoption and use of HIE across hospitals and other health care facilities in other states or regions with similar settings. Therefore, using the most recent available survey data, we aimed to explore changes in the proportion of HIE participation and associated factors between 2014 and 2020 and barriers to interoperability as of 2020 across SC hospitals to inform policy makers and providers on how to enact additional policy

interventions to ensure HIE adoption and use for a patient-centric health care system.

Methods

Study Design

We conducted a retrospective, longitudinal analysis of 69 individual hospitals in SC using the American Hospital Association (AHA) Annual Surveys IT Supplement from 2014 to 2020. This survey was sent to the chief executive officer of every hospital in the United States for completion by themselves or by the most knowledgeable person in each hospital. The AHA Annual Survey included questions about organization structure, technology adoption, and other topics about service provision. The AHA Annual Survey IT Supplement database collects data on facility-level adoption and implementation of the US Department of Health and Human Services Promoting Interoperability initiative, including computerized system capabilities, patient engagement, HIE, barriers to HIE and interoperability, and other factors [31]. The response rates of the AHA IT survey among SC hospitals ranged from 44.9% in 2014 to 59.4% in 2020. The primary question was whether a hospital participated in local HIE activities in a year. According to the AHA Annual Survey, HIE involvement was assessed by indicating the level of participation in a state, regional, and/or local HIE or HIO. Using historical responses, we were able to carry forward and impute the missing values of 14 hospitals based on the following scenarios:

1. Hospitals that responded as having participated in HIE in the previous and subsequent years but did not respond to the survey in a given year were coded as participants.
2. Hospitals reporting no participation in the previous and subsequent years but did not respond to the survey in a given year were coded as nonparticipants.
3. Hospitals that reported no local HIE/HIO or no participation in HIE/HIO in a given year but did not respond in the previous year were coded as nonparticipants in the previous year. For example, if a hospital did not participate in HIE in 2015 and did not report whether they participated in 2014, the hospital would be coded as nonparticipants in 2014.
4. Hospitals that reported having participated in HIE/HIO in the previous year but did not respond in a given year were coded as nonparticipants.

The final analytic data set included 69 unique hospitals, but varied by year, ranging from 56 hospitals in 2014 to 66 hospitals in 2020. In 2014, of the 56 hospitals, only 27 responded on barriers to send or receive patient health information and other barriers related to exchanging patient health information. In 2020, of the 66 hospitals, 40 responded on barriers to send or receive patient health information and other barriers related to exchanging patient health information.

Variables

The primary independent variable was hospital location by rurality, categorized by Rural-Urban Commuting Area codes into urban (primary codes: 1-3) or rural (4-10). We also derived the following factors from the AHA Annual Surveys: number of beds staffed (1, <100 beds; 2, 100-299 beds; 3, >300 beds), teaching status, ownership (public nonfederal, private for-profit, or private nonprofit hospitals), system-affiliated hospitals (yes or no), primary services code (general or specialty hospitals), Medicare payment scheme (prospective payment system, CAHs, rural referral hospitals or sole community hospitals), ratio of Medicare inpatient days to total inpatient days, and ratio of Medicaid inpatient days to total inpatient days. Furthermore, we set survey year as covariate to examine annual linear trends in the proportion of HIE adoption across SC hospitals.

The second outcome variable was barriers to send, receive, or find patient health information to or from other organizations or hospital systems, which was not imputed because the barriers can change over time. To capture the main barriers to exchange and interoperability of patient health information, each hospital was asked, “Which of the following issues has your hospital experienced when trying to electronically (not eFax) send, receive or find (query) patient health information to/from other organizations or hospital systems?” Per the Healthcare Information and Management System Society, interoperability is the ability of different information systems, devices, and applications to access, exchange, integrate, and cooperatively use data, while exchange is the ability to send or to receive information but not necessary to integrate and harmonize data for further use [32].

Statistical Analysis

We first compared hospital characteristics by the status of HIE participation and presented geographic distributions of HIE participation across SC hospitals in 2014 and 2020. We then conducted chi-square tests to compare hospital characteristics by HIE participation status. Cross-sectional multivariable logistic regression models clustered at the hospital level to account for repeated observations and weighted by bed size were used to identify factors associated with HIE participation across hospitals. Additionally, we calculated the frequency of hospitals reporting each barrier related to electronically sending or receiving patient health information or other financial or technical barriers to exchanging patient health information. All analyses were conducted using Stata 15.0 (StataCorp LLC).

Ethical Considerations

The University of South Carolina Institutional Review Board exempted this study protocol.

Results

Proportion of Hospitals With HIE Technologies

In 2014, 24 of 56 (43%) hospitals reported participating in HIE. Private for-profit hospitals, nonteaching hospitals, small hospitals (<100 beds), non–system-affiliated hospitals, and specialty hospitals were less likely to participate in HIE than their peers. In 2020, 54 of 66 (82%) hospitals reported participating in HIE (Table 1). In 2020, 77% (13/17) of rural hospitals reported participating in HIE, which was comparable to the 84% (41/49) participation rate among urban hospitals. Between 2014 and 2020, there was an increasing trend in HIE participation among sample hospitals (Figure 1). The geographic distribution of these hospitals per their self-reported HIE

participation status is illustrated in Figure 2. In 2020, hospitals in the Upstate region mostly participated in HIE or HIO, whereas hospitals in many rural counties did not participate despite the HIE network availability in their local areas. The proportions of rural and urban hospitals with HIE or HIO participation increased at similar rates between 2014 to 2020 (Multimedia Appendix 1).

Table 1. Hospital characteristics by health information exchange participation status in 2014 and 2020.[a]

Characteristics	2014 (n=56)			2020 (n=66)		
	Yes, n (%)	No, n (%)	*P* value	Yes, n (%)	No, n (%)	*P* value
All respondent hospitals	24 (43)	32 (57)	N/A[b]	54 (82)	12 (19)	N/A
Rurality			.10			.51
Rural	3 (23)	10 (77)		13 (77)	4 (24)	
Urban	21 (49)	22 (51)		41 (84)	8 (16)	
Hospital ownership			.37			.19
Public nonfederal	7 (39)	11 (61)		16 (73)	6 (27)	
Private nonprofit	11 (55)	9 (45)		24 (92)	2 (8)	
Private for-profit	6 (33)	12 (67)		14 (78)	4 (22)	
Teaching hospitals			.02			.43
Yes	3 (100)	0 (0)		3 (100)	0 (0)	
No	21 (40)	32 (60)		51 (81)	12 (19)	
Hospital bed size			.02			.04
<100 beds	8 (28)	21 (72)		34 (90)	4 (11)	
100-299 beds	7 (47)	8 (53)		9 (60)	6 (40)	
>300 beds	9 (75)	3 (25)		11 (85)	2 (15)	
Affiliation to health system			.12			.06
Yes	19 (49)	20 (51)		45 (87)	7 (14)	
No	5 (29)	12 (71)		9 (64)	5 (36)	
Specialty hospitals			.06			.95
Yes	3 (21)	11 (79)		14 (82)	3 (18)	
No	21 (50)	21 (50)		40 (82)	9 (18)	
Medicare payment scheme			.05			.88
Critical access hospitals	0 (0)	4 (100)		3 (75)	1 (25)	
Rural referral hospitals	2 (100)	0 (0)		6 (86)	1 (14)	
Sole community hospitals	1 (50)	1 (50)		3 (75)	1 (25)	
Ratio of Medicare inpatient days to total inpatient days			.57			.10
<47.0%	8 (38)	13 (62)		17 (74)	6 (26)	
47.0%-58.8%	10 (53)	9 (47)		15 (75)	5 (25)	
>58.8%	6 (38)	10 (63)		22 (96)	1 (4)	
Ratio of Medicaid inpatient days to total inpatient days			.21			.79
<8.0%	6 (30)	14 (70)		15 (79)	4 (21)	
8.0%-14.4%	11 (58)	8 (42)		19 (86)	3 (14)	
>14.4%	7 (41)	10 (59)		20 (80)	5 (20)	

[a]Data are presented with consideration of missing values for the category variables.

[b]N/A: not applicable.

Figure 1. Proportion of hospitals with health information exchange technologies from 2014 to 2020. The proportion of hospitals participating in a health information exchange HIE or HIO increased over time (P<.001). Results stratified by hospital location can be found in Multimedia Appendix 1. HIE: health information exchange; HIO: health information organization.

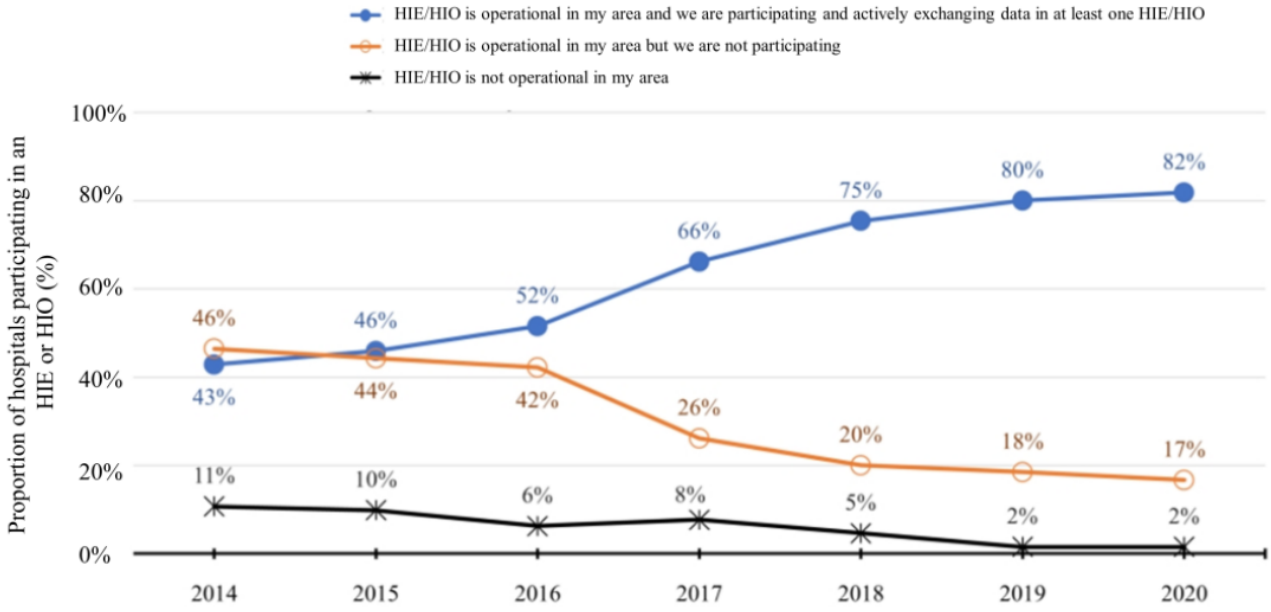
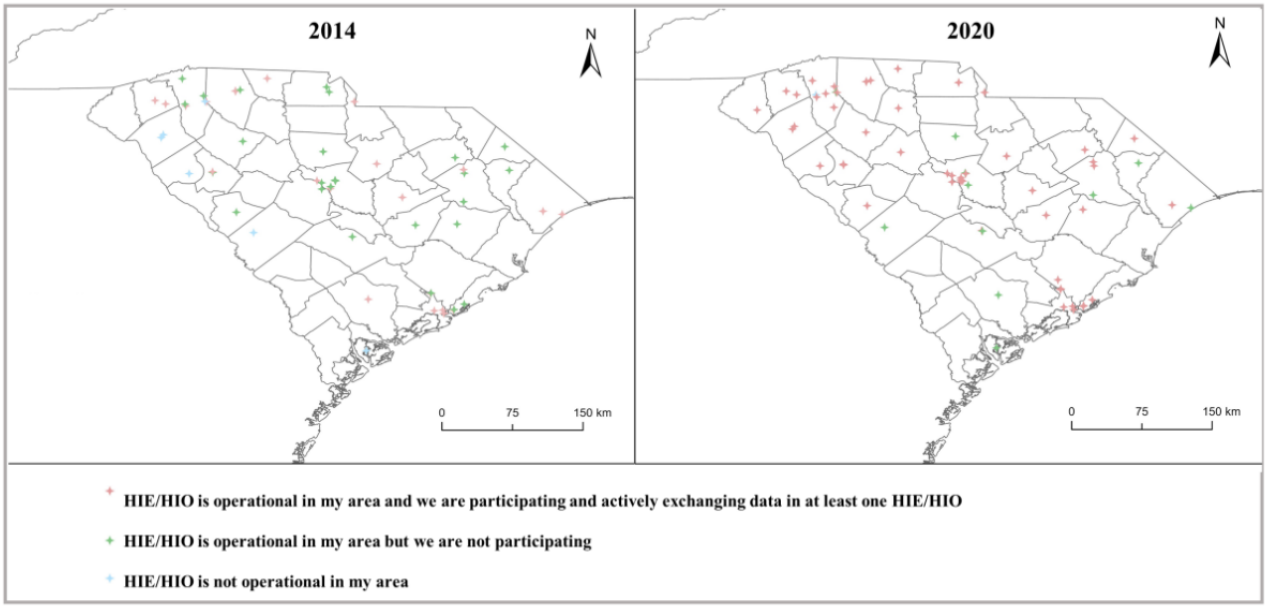


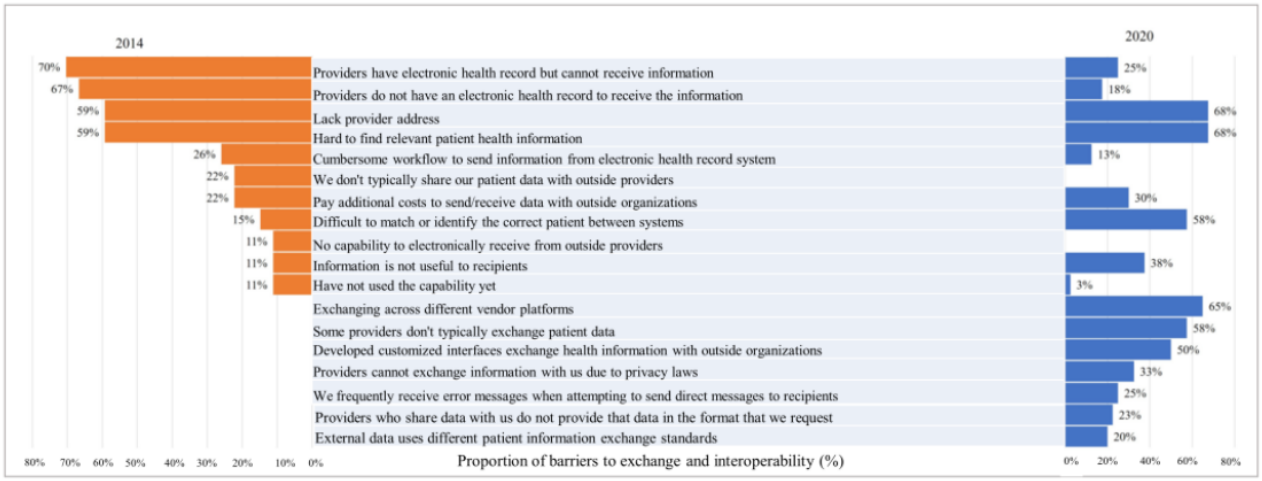
Figure 2. Geographic distribution of HIE participation across South Carolina hospitals in 2014 and 2020. HIE: health information exchange; HIO: health information organization.



Barriers to Exchange and Interoperability Across SC Hospitals

As shown in Figure 3, among hospital respondents, the leading barriers to HIE in 2020 were difficulties in locating the contact information of the providers to send patient health information (27/40, 68%) and finding relevant patient health information (27/40, 68%), followed by several technical and systemic issues, including exchanging across different vendor platforms (26/40, 65%), difficulty matching or identifying the correct patient between systems (23/40, 58%), and providers that do not typically exchange patient data (23/40, 58%). Additionally, 33% (13/40) of sample hospitals reported that providers could not exchange information due to privacy laws as barriers to exchange and interoperability. In 2014, the most commonly reported barriers to exchange and interoperability were providers having EHRs but being unable to receive information (19/27, 70%), providers lacking an EHR to receive information (18/27, 67%), difficulty locating providers' address (16/27, 59%), difficulties in finding relevant patient health information (16/27, 59%), and cumbersome workflow to send information from the EHR system (7/27, 26%).

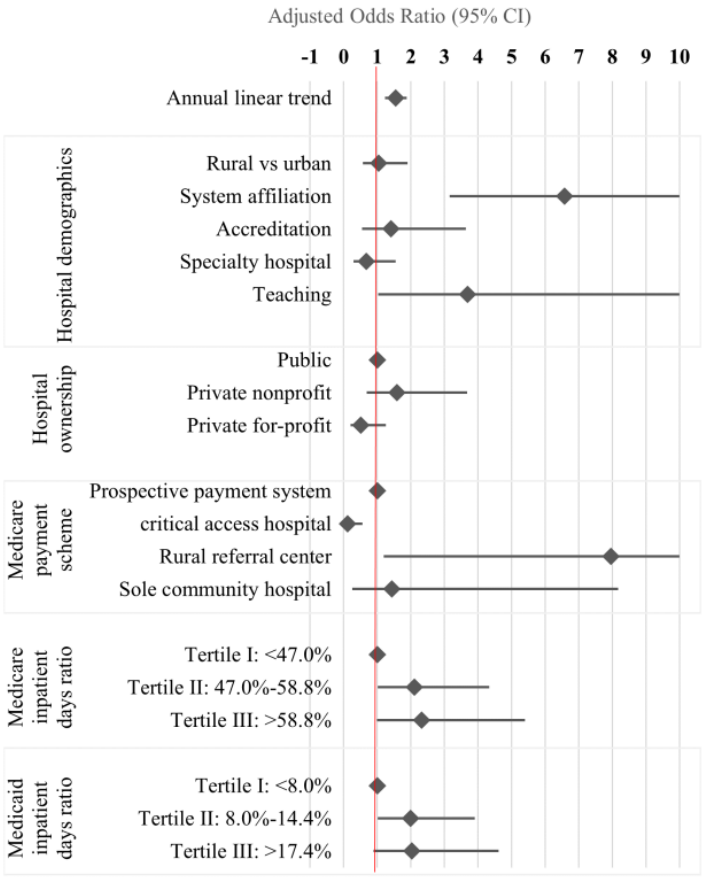
Figure 3. Barriers to exchange and interoperability across South Carolina hospitals in 2014 (n=27) and 2020 (n=40). EHR: electronic health record.



Factors Associated With Hospitals' HIE Participation

After controlling for other hospital factors, teaching hospitals (adjusted odds ratio [AOR] 3.7, 95% CI 1.0-13.3), rural referral hospitals (AOR 8.0, 95% CI 1.2-53.4), and system-affiliated hospitals (AOR 6.6, 95% CI 3.2-13.7) had higher odds of participating in HIE, whereas CAHs (AOR 0.1, 95% CI 0.02-0.6) were less likely to participate in HIE. Hospitals with greater ratios of Medicare inpatient days to total inpatient days or Medicaid inpatient days to total inpatient days also reported higher odds of HIE participation (Figure 4).

Figure 4. Multivariable analysis of factors associated with hospitals' participation in health information exchange from 2014 to 2020. Multivariable logistic regression models were used with adjusted odds ratios and 95% CIs calculated from standard errors clustered at the hospital level and weighted by hospital bed size.



Discussion

Principal Findings

In 2014, the Office of the National Coordinator for Health Information Technology launched a 10-year road map for the United States to achieve interoperability of EHRs by 2024. In this longitudinal observational study of 69 SC hospitals, we found that federal and state efforts to build an HIE system across all hospitals have been successful in SC, with over 80% of SC hospitals participating in HIE as of 2020; however, there is still a long way to go to achieve Meaningful Use Stage 3 for improved health outcomes. We also found teaching hospitals, system-affiliated hospitals, rural referral hospitals, and hospitals with greater ratios of Medicare or Medicaid inpatient days to total inpatient days were more likely to participate in HIE, while CAHs were less likely to participate in HIE. Yet, key barriers for effective use of HIE for patient care coordination and improved health outcomes remain, including technical, data management, and legal issues. These findings highlight the need to ensure the accuracy of patient information, standardization of health information across various systems, and training of privacy and security regulations to optimize HIE use for data exchange and patient care transition.

As of 2020, although over 80% of SC hospitals participated in HIE, our findings suggest that rural hospitals experienced no greater probability of HIE participation. This might be caused by multiple barriers to HIE implementation, including challenges with reporting, workflow, and technology capacity (eg, availability of adequate broadband), even with generous financial incentives [21]. Our findings also indicate that CAHs were less likely to participate in HIE than prospective payment system hospitals, leading to delayed development of EHR interfaces by private hospitals at the early stage of HIE expansion. This result is consistent with the findings of a previous study that CAHs with EHRs still struggle to keep up with other advanced functions, such as patient engagement and clinical data analytics, as of 2018 [19]; these hospitals face skyrocketing costs and a lack of technical expertise when developing HIE interfaces, making HIE adoption a prohibitive move [19,25]. Given that CAHs were less likely to adopt advanced functions for electronic records [33], challenges for CAHs to participate and use HIE may be further exacerbated. Moreover, hospitals with a greater ratio of Medicaid or Medicare inpatient days to total inpatient days reported higher probabilities of HIE participation. This result was consistent with prior research, indicating that broader changes in the payment of care may encourage health care providers to use EHRs [34].

Many barriers to HIE implementation and use exist. Despite the HIE functionality, we found that much critical information was missing or hard to find. For example, over two-thirds of HIE-active hospitals reported difficulties locating providers' addresses to send and receive information or finding relevant patient health information. Even with information availability, some had difficulties matching or identifying the correct patients between health care or EHR systems. On top of the information discontinuity, technical challenges in exchanging across different vendor platforms and the fact that some providers do not typically exchange patient data with outside providers further fuel the structural obstacles to the exchange and interoperability of patient health information across hospitals. In the era of telehealth and HIE across clinicians and hospitals for optimal quality of care [35,36], it is striking to find over half of SC hospitals reported barriers to the exchange of patient health information across different vendor platforms. Over half of providers do not typically exchange patient data, which may be because increasing data volumes and types is difficult and labor-intensive to match [37,38]. These barriers raise concerns of approaches at systematic and provider levels to improve the quality of health information, the health information system, and telehealth services.

Lacking provider addresses and difficulties in finding relevant patient health information were the main barriers to exchange in SC hospitals in 2014, suggesting that compatibility issues between EHRs and HIE systems are a significant systematic barrier [39]. Technical integration across distinct EHR platforms can be a real challenge given that Stage 1 of Meaningful Use did not require usability, data integrity, harmonization, and terminology mapping and matching among certified EHR systems. Moreover, the inability to get vendor support to address specific needs within hospitals limits the use of HIE for patient care transitions. Vendor issues are inextricably linked to EHR utility in many cases. Hospitals often need to use additional tools (eg, web-based data entry platforms) to retrieve specific data from their EHRs, limiting interoperability and optimization of HIE among hospitals.

Ensuring usability of information in clinical decision-making and the perception of accomplishing goals is critical for ensuring the sustainability of HIE use across hospitals facing barriers to exchange and interoperability [40]. However, in 2020, nearly 40% of hospitals reported that many recipients of electronic care summaries found that the information was not useful, and about 30% of providers could not exchange health information due to privacy laws or additional costs, which have posed obstacles to the use of HIE among health care providers. Given that physician-level variation in EHR documents can impede effective use of patient health information [41], the use of HIE may be lower when patients are unfamiliar with the health care provider [42]. This result might also be related to the fact that health care providers are more concerned about economic and competitive risks than perceived benefits [14,27]; therefore, they may complain about these issues [27], which might hinder the use of HIE in SC hospitals and discourage opportunities for improving patient care and outcomes [40]. Additionally, a previous study revealed that state and federal health information privacy laws, beyond economic and technical barriers, have been a significant obstacle to expanding HIE [27]. As the exchange of patient health information is a trust-related behavior, providers' perceived trust in HIE's technical capabilities, skills, and benefits warrants improvement [43]. Use of HIE by the "marquee user" should aim to attract more users to the platform and eliminate these barriers [16]. These findings suggest that the existing infrastructure may not consider the unique needs of patients who often access multiple health care service sites across multiple geographies. As many nontechnological factors may hinder the effective use of HIE

[44], more efforts beyond addressing workflow and technical issues are needed to make the HIE sustainable [45]. This result also raises concerns about the full connectivity of the statewide framework, including connecting unaffiliated ambulatory care practices [46] and skilled nursing facilities [47].

Limitations

This study has some limitations. First, the AHA Annual Surveys IT Supplement asked about hospital-wide use of HIE via a single item, and the hospitals did not report the extent of use. About 60% of hospitals responded to items on the barriers to exchange and interoperability. Assuming that SC hospitals without HIE tend not to respond to items on the barriers to exchange and interoperability, the current state-level barriers to exchange and interoperability may be underestimated. Second, we did not have data about local HIE needs. Third, the sample hospitals that reported actively participating in HIE may not have exercised HIE to its full potential. Fourth, the AHA Annual Surveys IT Supplement did not collect comprehensive vendor-specific information, and heterogeneity of vendors was not accounted for in this study. Nevertheless, our findings have revealed vendor-specific barriers to exchange and interoperability. Future research is warranted to examine how often HIE has been used across patients or patient visits/admissions and to assess key factors facilitating the optimal use of HIE.

Conclusions

Nearly all hospitals in SC participate in HIE. However, barriers to exchange and interoperability remain, including technical, data management, user training and support, and legal issues, highlighting a crucial missed opportunity to improve health outcomes. These findings imply the need to incentivize optimization of HIE and seamless patient information exchange across facilities by facilitating and implementing standardization of health information across various HIE systems and by training HIE stakeholders about privacy and security regulations to ensure smooth, safe, and secure patient care transitions. Future policies and efforts should promote collaborations with vendors to reduce platform compatibility issues and increase user engagement and technical training and support to facilitate effective, accurate, and efficient exchange of patient health information.

Acknowledgments

PH received funding to support this research from the South Carolina Health and Human Service Department.

Authors' Contributions

PH and ZL designed the study; PH performed the data analysis; all authors were involved the interpretation of the results; ZL and PH drafted the manuscript; and MAM, JME, and DW critically revised the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Proportion of rural and urban hospitals participating in a health information exchange or health information organization from 2014 to 2020.
[PNG File , 224 KB - medinform_v11i1e40959_app1.png]

References

1. Adler-Milstein J, McAfee AP, Bates DW, Jha AK. The state of regional health information organizations: current activities and financing. Health Aff (Millwood) 2008 Nov 11;27(1):w60-w69. [doi: 10.1377/hlthaff.27.1.w60] [Medline: 18073225]
2. Adler-Milstein J, DesRoches CM, Kralovec P, Foster G, Worzala C, Charles D, et al. Electronic health record adoption in US hospitals: progress continues, but challenges persist. Health Aff (Millwood) 2015 Dec 01;34(12):2174-2180. [doi: 10.1377/hlthaff.2015.0992] [Medline: 26561387]
3. Adjerid I, Acquisti A, Telang R, Padman R, Adler-Milstein J. The impact of privacy regulation and technology incentives: the case of health information exchanges. Manag Sci 2016 Apr 01;62(4):1042-1063 [FREE Full text] [doi: 10.1287/mnsc.2015.2194]
4. Blumenthal D. Launching HITECH. N Engl J Med 2010 Feb 04;362(5):382-385. [doi: 10.1056/NEJMp0912825] [Medline: 20042745]
5. Medicare and Medicaid programs; electronic health record incentive program--stage 2. Final rule. Fed Regist 2012 Sep 04;77(171):53967-54162 [FREE Full text] [Medline: 22946138]
6. Walker J, Pan E, Johnston D, Adler-Milstein J, Bates DW, Middleton B. The value of health care information exchange and interoperability. Health Aff (Millwood) 2005 Jun 05;Suppl Web Exclusives:W5-W10. [doi: 10.1377/hlthaff.w5.10] [Medline: 15659453]

7. Hripcsak G, Kaushal R, Johnson KB, Ash JS, Bates DW, Block R, et al. The United Hospital Fund meeting on evaluating health information exchange. J Biomed Inform 2007 Dec 01;40(6 Suppl):S3-S10 [FREE Full text] [doi: 10.1016/j.jbi.2007.08.002] [Medline: 17919986]
8. Chen M, Guo S, Tan X. Does health information exchange improve patient outcomes? Empirical evidence from Florida hospitals. Health Aff (Millwood) 2019 Feb 04;38(2):197-204. [doi: 10.1377/hlthaff.2018.05447] [Medline: 30715992]
9. Ryan AM, Krinsky S, Adler-Milstein J, Damberg CL, Maurer KA, Hollingsworth JM. Association between hospitals' engagement in value-based reforms and readmission reduction in the hospital readmission reduction program. JAMA Intern Med 2017 Jun 01;177(6):862-868 [FREE Full text] [doi: 10.1001/jamainternmed.2017.0518] [Medline: 28395006]
10. Interoperability and Patient Access Fact Sheet. CMS.gov. 2020 Mar 09. URL: https://www.cms.gov/newsroom/fact-sheets/interoperability-and-patient-access-fact-sheet [accessed 2022-07-11]
11. Shields MC, Ritter G, Busch AB. Electronic health information exchange at discharge from inpatient psychiatric care in acute care hospitals. Health Aff (Millwood) 2020 Jun 04;39(6):958-967 [FREE Full text] [doi: 10.1377/hlthaff.2019.00985] [Medline: 32479237]
12. Cohen GR, Adler-Milstein J. Meaningful use care coordination criteria: perceived barriers and benefits among primary care providers. J Am Med Inform Assoc 2016 Apr 01;23(e1):e146-e151 [FREE Full text] [doi: 10.1093/jamia/ocv147] [Medline: 26567327]
13. Pylypchuk Y, Johnson C, Patel V. State of interoperability among U.S. non-federal acute care hospitals in 2018. HealthIT.gov. 2020 May 01. URL: https://www.healthit.gov/data/data-briefs/state-interoperability-among-us-non-federal-acute-care-hospitals-2018 [accessed 2023-02-15]
14. Adler-Milstein J, Jha AK. Health information exchange among U.S. hospitals: who's in, who's out, and why? Healthc (Amst) 2014 Mar 01;2(1):26-32. [doi: 10.1016/j.hjdsi.2013.12.005] [Medline: 26250086]
15. Lin SC, Everson J, Adler-Milstein J. Technology, incentives, or both? Factors related to level of hospital health information exchange. Health Serv Res 2018 Feb 28;53(5):3285-3308 [FREE Full text] [doi: 10.1111/1475-6773.12838] [Medline: 29492959]
16. Miller AR, Tucker C. Health information exchange, system size and information silos. J Health Econ 2014 Jan 01;33:28-42. [doi: 10.1016/j.jhealeco.2013.10.004] [Medline: 24246484]
17. Everson J, Adler-Milstein J. Gaps in health information exchange between hospitals that treat many shared patients. J Am Med Inform Assoc 2018 Sep 01;25(9):1114-1121 [FREE Full text] [doi: 10.1093/jamia/ocy089] [Medline: 30010887]
18. Thorn SA, Carter MA, Bailey JE. Emergency physicians' perspectives on their use of health information exchange. Ann Emerg Med 2014 Mar 01;63(3):329-337. [doi: 10.1016/j.annemergmed.2013.09.024] [Medline: 24161840]
19. Apathy NC, Holmgren AJ, Adler-Milstein J. A decade post-HITECH: critical access hospitals have electronic health records but struggle to keep up with other advanced functions. J Am Med Inform Assoc 2021 Aug 13;28(9):1947-1954 [FREE Full text] [doi: 10.1093/jamia/ocab102] [Medline: 34198342]
20. Furukawa MF, King J, Patel V, Hsiao C, Adler-Milstein J, Jha AK. Despite substantial progress in EHR adoption, health information exchange and patient engagement remain low in office settings. Health Aff (Millwood) 2014 Sep 04;33(9):1672-1679. [doi: 10.1377/hlthaff.2014.0445] [Medline: 25104827]
21. Heisey-Grove DM. Variation in rural health information technology adoption and use. Health Aff (Millwood) 2016 Feb 04;35(2):365-370. [doi: 10.1377/hlthaff.2015.0861] [Medline: 26791835]
22. Krakow M, Hesse BW, Oh A, Patel V, Vanderpool RC, Jacobsen PB. Addressing rural geographic disparities through health IT: initial findings from the health information national trends survey. Med Care 2019 Jun 01;57 Suppl 6 Suppl 2:S127-S132. [doi: 10.1097/MLR.0000000000001028] [Medline: 31095051]
23. Chen J, Amaize A, Barath D. Evaluating telehealth adoption and related barriers among hospitals located in rural and urban Aaeas. J Rural Health 2021 Sep 01;37(4):801-811 [FREE Full text] [doi: 10.1111/jrh.12534] [Medline: 33180363]
24. Kruse CS, Regier V, Rheinboldt KT. Barriers over time to full implementation of health information exchange in the United States. JMIR Med Inform 2014 Sep 30;2(2):e26 [FREE Full text] [doi: 10.2196/medinform.3625] [Medline: 25600635]
25. Yeager VA, Walker D, Cole E, Mora AM, Diana ML. Factors related to health information exchange participation and use. J Med Syst 2014 Aug 01;38(8):1-9. [doi: 10.1007/s10916-014-0078-1] [Medline: 24957395]
26. Lee L, Whitcomb K, Galbreth M, Patterson D. A strong state role in the HIE. Lessons from the South Carolina Health Information Exchange. J AHIMA 2010 Jun 01;81(6):46-50. [Medline: 20614703]
27. Mello MM, Adler-Milstein J, Ding KL, Savage L. Legal barriers to the growth of health information exchange-boulders or pebbles? Milbank Q 2018 Mar 05;96(1):110-143 [FREE Full text] [doi: 10.1111/1468-0009.12313] [Medline: 29504197]
28. Patel V, Abramson EL, Edwards A, Malhotra S, Kaushal R. Physicians' potential use and preferences related to health information exchange. Int J Med Inform 2011 Mar 01;80(3):171-180. [doi: 10.1016/j.ijmedinf.2010.11.008] [Medline: 21156351]
29. Pevnick JM, Claver M, Dobalian A, Asch SM, Stutman HR, Tomines A, et al. Provider stakeholders' perceived benefit from a nascent health information exchange: a qualitative analysis. J Med Syst 2012 Apr 22;36(2):601-613 [FREE Full text] [doi: 10.1007/s10916-010-9524-x] [Medline: 20703673]
30. Bronsoler A, Doyle J, Schmit C, Van Reenen J. The role of state policy in fostering health information exchange in the United States. NEJM Catalyst 2022 Dec 21;4(1) [FREE Full text] [doi: 10.1056/cat.22.0302]

31. 2020 AHA annual survey information technology supplement: public health and COVID-19 focus. American Hospital Association. 2020 Dec 01. URL: https://www.ahadata.com/system/files/media/file/2021/12/2020_AHAIT_Survey-Dec092021_0.pdf [accessed 2022-12-26]
32. Interoperability in healthcare. Healthcare Information and Management Systems Society. URL: https://www.himss.org/resources/interoperability-healthcare [accessed 2023-08-20]
33. Adler-Milstein J, Holmgren AJ, Kralovec P, Worzala C, Searcy T, Patel V. Electronic health record adoption in US hospitals: the emergence of a digital "advanced use" divide. J Am Med Inform Assoc 2017 Nov 01;24(6):1142-1148 [FREE Full text] [doi: 10.1093/jamia/ocx080] [Medline: 29016973]
34. Adler-Milstein J, Salzberg C, Franz C, Orav EJ, Newhouse JP, Bates DW. Effect of electronic health records on health care costs: longitudinal comparative evidence from community practices. Ann Intern Med 2013 Jul 16;159(2):97-104. [doi: 10.7326/0003-4819-159-2-201307160-00004] [Medline: 23856682]
35. Williams C, Mostashari F, Mertz K, Hogin E, Atwal P. From the Office of the National Coordinator: the strategy for advancing the exchange of health information. Health Aff (Millwood) 2012 Mar 04;31(3):527-536. [doi: 10.1377/hlthaff.2011.1314] [Medline: 22392663]
36. Finkelstein J, Barr MS, Kothari PP, Nace DK, Quinn M. Patient-centered medical home cyberinfrastructure current and future landscape. Am J Prev Med 2011 May 01;40(5 Suppl 2):S225-S233. [doi: 10.1016/j.amepre.2011.01.003] [Medline: 21521598]
37. Luxton DD, Kayl RA, Mishkind MC. mHealth data security: the need for HIPAA-compliant standardization. Telemed J E Health 2012 May 08;18(4):284-288. [doi: 10.1089/tmj.2011.0180] [Medline: 22400974]
38. Norton JM, Ip A, Ruggiano N, Abidogun T, Camara DS, Fu H, et al. Assessing progress toward the vision of a comprehensive, shared electronic care plan: scoping review. J Med Internet Res 2022 Jun 10;24(6):e36569 [FREE Full text] [doi: 10.2196/36569] [Medline: 35687382]
39. Ranade-Kharkar P, Pollock SE, Mann DK, Thornton SN. Improving clinical data integrity by using data adjudication techniques for data received through a health information exchange (HIE). AMIA Annu Symp Proc 2014 Nov 14;2014:1894-1901 [FREE Full text] [Medline: 25954462]
40. Holmgren AJ, Patel V, Adler-Milstein J. Progress in interoperability: measuring US hospitals' engagement in sharing patient data. Health Aff (Millwood) 2017 Oct 01;36(10):1820-1827. [doi: 10.1377/hlthaff.2017.0546] [Medline: 28971929]
41. Cohen GR, Friedman CP, Ryan AM, Richardson CR, Adler-Milstein J. Variation in physicians' electronic health record documentation and potential patient harm from that variation. J Gen Intern Med 2019 Nov 10;34(11):2355-2367 [FREE Full text] [doi: 10.1007/s11606-019-05025-3] [Medline: 31183688]
42. Vest JR, Zhao H, Jasperson J, Gamm LD, Ohsfeldt RL. Factors motivating and affecting health information exchange usage. J Am Med Inform Assoc 2011 Jan 24;18(2):143-149 [FREE Full text] [doi: 10.1136/jamia.2010.004812] [Medline: 21262919]
43. Esmaeilzadeh P. The impacts of the perceived transparency of privacy policies and trust in providers for building trust in health information exchange: empirical study. JMIR Med Inform 2019 Nov 26;7(4):e14050 [FREE Full text] [doi: 10.2196/14050] [Medline: 31769757]
44. Vest JR. More than just a question of technology: factors related to hospitals' adoption and implementation of health information exchange. Int J Med Inform 2010 Dec 01;79(12):797-806. [doi: 10.1016/j.ijmedinf.2010.09.003] [Medline: 20889370]
45. Rudin RS, Motala A, Goldzweig CL, Shekelle PG. Usage and effect of health information exchange. Ann Intern Med 2014 Dec 02;161(11):803. [doi: 10.7326/m14-0877]
46. Posnack S. Connecting health and care for the nation: a shared nationwide interoperability roadmap. HealthIT.gov. URL: https://ncvhs.hhs.gov/wp-content/uploads/2015/10/Day-2-NCVHS-Dept-Update-POSNACK.pdf [accessed 2022-07-11]
47. Adler-Milstein J, Raphael K, O'Malley TA, Cross DA. Information sharing practices between US hospitals and skilled nursing facilities to support care transitions. JAMA Netw Open 2021 Jan 04;4(1):1-13 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.33980] [Medline: 33443582]

Abbreviations

AOR: adjusted odds ratio
AHA: American Hospital Association
CAH: Critical Access Hospital
EHR: electronic health record
HIE: health information exchange
HIO: health information organization
HITECH: Health Information Technology for Economic and Clinical Health
SC: South Carolina

Edited by J Klann; submitted 11.07.22; peer-reviewed by M Esdar, D Chrimes; comments to author 17.12.22; revised version received 15.02.23; accepted 29.08.23; published 28.09.23.

Please cite as:
Li Z, Merrell MA, Eberth JM, Wu D, Hung P
Successes and Barriers of Health Information Exchange Participation Across Hospitals in South Carolina From 2014 to 2020: Longitudinal Observational Study
JMIR Med Inform 2023;11:e40959
URL: https://medinform.jmir.org/2023/1/e40959
doi:10.2196/40959
PMID:37768730

Original Paper

Perspectives on Challenges and Opportunities for Interoperability: Findings From Key Informant Interviews With Stakeholders in Ohio

Daniel M Walker[1,2], MPH, PhD; Willi L Tarver[2,3], MLiS, DrPH; Pallavi Jonnalagadda[2], MBBS, DrPH; Lorin Ranbom[4], BA; Eric W Ford[5], MPH, PhD; Saurabh Rahurkar[2,6], BDS, DrPH

[1]Department of Family and Community Medicine, College of Medicine, The Ohio State University, Columbus, OH, United States

[2]The Center for the Advancement of Team Science, Analytics, and Systems Thinking, College of Medicine, The Ohio State University, Columbus, OH, United States

[3]Department of Internal Medicine, College of Medicine, The Ohio State University, Columbus, OH, United States

[4]Government Resource Center, College of Medicine, The Ohio State University, Columbus, OH, United States

[5]Department of Healthcare Organization and Policy, School of Public Health, University of Alabama, Birmingham, AL, United States

[6]Department of Biomedical Informatics, College of Medicine, The Ohio State University, Columbus, OH, United States

Corresponding Author:
Daniel M Walker, MPH, PhD
Department of Family and Community Medicine
College of Medicine
The Ohio State University
Suite 5000
700 Ackerman Rd
Columbus, OH, 43202
United States
Phone: 1 203 988 1800
Email: daniel.walker@osumc.edu

Abstract

Background: Interoperability—the exchange and integration of data across the health care system—remains a challenge despite ongoing policy efforts aimed at promoting interoperability.

Objective: This study aimed to identify current challenges and opportunities to advancing interoperability across stakeholders.

Methods: Primary data were collected through qualitative, semistructured interviews with stakeholders (n=24) in Ohio from July to October 2021. Interviewees were sampled using a stratified purposive sample of key informants from 4 representative groups as follows: acute care and children's hospital leaders, primary care providers, behavioral health providers, and regional health information exchange networks. Interviews focused on key informant perspectives on electronic health record implementation, the alignment of public policy with organizational strategy, interoperability implementation challenges, and opportunities for health information technology. The interviews were transcribed verbatim followed by rigorous qualitative analysis using directed content analysis.

Results: The findings illuminate themes related to challenges and opportunities for interoperability that align with technological (ie, implementation challenges, mismatches in interoperability capabilities across stakeholders, and opportunities to leverage new technology and integrate social determinants of health data), organizational (ie, facilitators of interoperability and strategic alignment of participation in value-based payment programs with interoperability), and environmental (ie, policy) domains.

Conclusions: Interoperability, although technically feasible for most providers, remains challenging for technological, organizational, and environmental reasons. Our findings suggest that the incorporation of end user considerations into health information technology development, implementation, policy, and standard deployment may support interoperability advancement.

(JMIR Med Inform 2023;11:e43848) doi:10.2196/43848

KEYWORDS

interoperability; health information exchange; health information technology; electronic health record; usability

Introduction

Background

Starting with the Health Information Technology for Economic and Clinical Health (HITECH) Act in 2009, the United States has invested over US$ 36 billion to promote interoperability—the ability of 2 or more systems to exchange and use information [1-3]—through health information exchange (HIE) networks and electronic health records (EHRs). HITECH promoted the adoption and implementation of certified EHRs by providing financial incentives through its "meaningful use" programs and funded grants that helped establish regional HIE networks [4]. Some of these financial incentive programs, such as the state Medicaid Provider Incentive Program (MPIP) were phased out in 2021. However, subsequent legislation such as the Patient Protection and Affordable Care Act reinforced this financing by advancing payment and care delivery models that use risk-based contracts to incentivize quality of care and patient outcomes, such as accountable care organizations, value-based care [5], and patient-centered medical homes [6], which stand to benefit from enhanced electronic data exchange. Other population health policy programs, including Comprehensive Primary Care (CPC) [7], CPC-Plus [8], and Primary Care First [9], further rely on robust data exchange among regional providers. Ongoing federal policy development has continued and enhanced support for meaningful use programs through the Medicare Access and Children's Health Insurance Program Reauthorization Act of 2015, leading to the Promoting Interoperability Program in 2018.

Recent reports show that 96% of acute care hospitals and 80% of primary care providers (PCPs) have implemented certified EHRs with interoperability capabilities [10,11]. However, actual use of data from an HIE network in clinical care encounters remains low [12]. Researchers have identified several barriers to the use of HIE networks, including poor user interfaces and lack of leadership support [13-15]. Although interoperable health information technology (HIT) is theorized to address these barriers, it continues to elude the health care system [16,17]. For instance, a recent study found that only 45% of all US hospitals engaged in the 4 core elements of interoperability, such as the capability of different EHR systems to find, send, receive, and use or integrate clinical information with one another [16]. Further, rural and smaller hospitals, ambulatory practices, as well as those ineligible for meaningful use incentives (eg, rehabilitation, long-term care, or behavioral health providers [BHPs]) lag behind large, integrated systems in adopting interoperability [18,19].

To promote interoperability, the 21st Century Cures Act (21CCA; 21CCA 2016) mandated the sharing of certain data elements, placed restrictions on information blocking, and promoted the use of application programming interfaces (APIs; eg, Fast Healthcare Interoperability Resources [FHIR]). The 21CCA also established the Trusted Exchange Framework Common Agreement (TEFCA) that provides an infrastructure model and governing approach for HIE networks [20]. However, in the sixth national survey of HIEs, 56% (84/151) of the regional HIEs planned to participate in TEFCA [21], which points to ongoing challenges that continue to limit data exchange. HIEs that intended to participate in TEFCA already had connections with HIEs in other states and participated in at least one national network. Therefore, the decision to participate in TEFCA may be determined by the alignment of HIE processes with existing data sharing rather than convincing HIEs of the benefits of participation [21].

Purpose

In alignment with policy efforts geared toward the promotion of interoperability, this study aimed to explore provider perspectives on the current state of interoperability challenges. Given the confluence of timing around the conclusion of state MPIPs and the ramp up of 21CCA and TEFCA, questions remain about the progress and remaining challenges related to achieving an interoperable health system. Moreover, previous research on interoperability typically focuses on a single perspective rather than those of multiple stakeholders. This multistakeholder lens is particularly relevant to consider when examining interoperability, as a key goal of this innovation is to connect disparate parts of the health system. Specifically, we conducted semistructured interviews with a stratified sample of key informants, including providers and individuals in leadership positions representing diverse organizations in Ohio, to identify barriers and facilitators to interoperability. The study's findings add to the body of knowledge about interoperability and may contribute to the efforts of state agencies and federal policy makers, such as the Office of the National Coordinator for Health Information Technology and the Centers for Medicare and Medicaid Services, to advance interoperability. For instance, our findings may provide evidence that supports alignment between the Health Insurance Portability and Accountability Act (HIPAA) with HITECH and 21CCA. Finally, technology vendors may benefit from increased understanding of the end user perspective of their applications to develop user-friendly software.

Methods

Study Setting and Design

We used a cross-sectional qualitative design to solicit multistakeholder perspectives on the state of interoperability in Ohio. Ohio provided financial incentives for the adoption of interoperable EHRs for eligible professionals and hospitals through the MPIP. Eligible professionals included those with an active Ohio Medicaid Provider Agreement, such as physicians, optometrists, dentists, certified nurse-midwives, nurse practitioners, and physician assistants practicing in a federally qualified health center or a rural health center led by a physician assistant. Eligible hospitals were also required to have an active Ohio Medicaid Provider Agreement and include acute care hospitals, critical access hospitals, cancer hospitals, and children's hospitals. Ineligible providers included most behavioral and mental health, long-term care, and home health providers. MPIP operated through 7 incentive cycles, with the final cycle occurring in 2021. On the basis of the 2021 Ohio Medicaid Electronic Health Records Survey for Practices and Hospitals [22], 23.87% (5593/23,435) of the eligible providers and hospitals had received at least one MPIP payment, 96.06%

(5280/5593) of the MPIP recipient providers and hospitals had adopted and used an EHR, and 90.73% (16,188/17,842) of those who did not receive an MPIP payment had adopted and used an EHR. Among the 735 ineligible providers, 72.1% (530/735) reported adopting and using an EHR. Epic Systems Corporation is the most prevalent EHR vendor in the state, with 36.80% (2058/5593) of MPIP recipient providers and hospitals and 56.56% (10,092/17,842) of non-MPIP recipient providers and hospitals using an Epic EHR. However, though Epic remained the EHR of choice for group practices (across multiple or single sites) and hospitals, individual practices were more likely to choose other EHR vendors (eg, NextGen and eClinicalWorks). Among the ineligible providers, there was substantial variation in EHR vendors, and CareLogic was the single most prevalent EHR vendor, with 7.8% (57/735) of providers adopting it. Presently, Ohio has 2 large-scale regional HIEs that facilitate electronic exchange of patient data. Almost 31.64% (7416/23,435) of MPIP eligible providers and hospitals had existing arrangements with the regional HIEs to share electronic patient-level clinical data, whereas 21.5% (158/735) of MPIP ineligible providers participated in the regional HIEs.

Data collection and analysis were guided by the technology-organization-environment (TOE) framework [23]. TOE is an organization-level theory that has been applied to explain how the 3 interacting contextual domains (ie, technology-organization-environment) influence a health care organization's technology-related decision-making [24].

Ethics Approval

This study was considered to have minimal risk and was approved by the Ohio State University institutional review board (2021B0378).

Sample Selection

We used a stratified purposive sampling approach to gather diverse perspectives on interoperability based on four representative groups: (1) acute care and children's hospital leaders, (2) PCPs, (3) BHPs (ie, providers or organizations that provide care for mental health, substance use disorders, stress-related physical symptoms, and life stressors and crises), and (4) regional HIEs (ie, organizations that facilitate information exchange within a network of facilities within a geographic boundary). In addition, both rural and urban subsamples within each of the 3 provider groups were interviewed to ensure geographic representation. We focused on key informants with first-hand perspectives on HIT adoption and its future directions. These key informants were administrative leaders with clinical and nonclinical backgrounds within organizations with decision-making capacity regarding HIT (eg, chief medical information officers, executive directors, chief executive officers, and strategy officers). The study leads received a list of potential key informants from the Ohio Department of Medicaid. Emails were sent to gauge interest in participation; key informants who agreed to participate (20 of 38 organizations invited) were then interviewed. To eliminate any potential conflict of interest, the Ohio Department of Medicaid was not notified of who agreed or refused to participate, and did not participate in any interviews.

Data Collection

We conducted semistructured interviews with key informants from July to October 2021. The interview guide was developed to ask about EHR implementation in general, the alignment of public policy with organizational strategy, interoperability implementation challenges, and opportunities to improve HIT effectiveness. The interview guide was piloted with administrative leaders (n=2; eg, chief information officer) at an urban hospital in Ohio. This process yielded two versions of the semistructured interview guide for (1) providers (ie, those from hospitals, primary care, and behavioral health) and (2) representatives of HIEs (Multimedia Appendix 1). All interviews were conducted remotely via Zoom and audio recorded.

Data Analysis

Interviews were transcribed verbatim and analyzed using directed content analysis—an approach that begins with the a priori codes from the TOE framework yet is permissive of emergent themes [25]. The coding team (DMW, WLT, and LP) met weekly to discuss the interviews and preliminary findings throughout the data collection phase. A preliminary codebook was developed by reviewing transcripts and identifying broad themes that emerged from the interview transcripts to organize the data into the 3 TOE domains. Next, to build consensus on the coding guide, the analysis team collectively reviewed common transcripts (n=3) to compare results, refine the codebook, and reconcile any coding discrepancies. The remaining transcripts were divided by organization type for coding. The team continued to compare findings throughout the analysis phase to achieve thematic saturation. The trustworthiness of our findings is ensured by our rigorous and iterative approach to analysis [26]. NVivo software (version 12) was used to support data coding and analysis.

Results

Overview

Overall, 24 key informants were interviewed representing 20 distinct organizations: 5 hospitals, 7 PCPs, 6 BHPs, and 2 HIEs. Of the 20 organizations, 15 (75%) were located in urban areas (Table 1).

Below, we report on the key themes and subthemes of the analysis organized by the domains of the TOE framework, including comparing and contrasting those that cut across organizational types as well as those specific to each individual organization.

Table 1. Key informant interview sample characteristics[a].

Organization type	Interviewees (n=24), n (%)	Organizations (n=20), n (%)	Urban area organizations (n=15), n (%)
Acute care and children's hospital	5 (21)	5 (25)	4 (27)
Primary care provider	7 (29)	7 (35)	5 (33)
Behavioral health provider	10 (42)	6 (30)	4 (27)
Regional health information exchange	2 (8)	2 (10)	2 (13)

[a]Some interviews (n=3) included multiple key informants.

Technology Domain

Overview

Three themes within the technology domain that related to the usability and technological aspects of interoperability were identified: (1) implementation challenges, (2) interoperability capabilities, and (3) opportunities (Table 2).

Table 2. Themes and primary subthemes in the technology domain.

Theme and primary subtheme	Representative quote
Implementation challenges	
Maintaining growing number of applications	"There's been an explosion of health IT [information technology] applications, vendors and products over the past decade, many of whom overlap and functionality intersect in ways that don't really allow for great interoperability, so just the challenge of sort of how do you meet all the needs, using all the various products out there, and still have a cohesive, reliable, safe experience is a challenge." [Hospital representative #15]
Integrating diverse sources of data into unified medical record	"Our community health centers are on a lot of different EHR [electronic health record] platforms, and those platforms don't talk to each other, and so the interoperability that we all desire is still not really there. So, we...are utilizing a health population, a population health tool that can sit over any EHR [electronic health record] platform, and that is allowing us to get some of the data aggregated, in spite of the lack of interoperability and communication between different EHRs." [Primary care provider #22]
Interoperability capabilities	
Connecting to state-hosted registries and databases (ie, state immunization registry)	"Every hospital has this issue, is that there's not really a good way to leverage the data being collected by the state vital statistic[s] for our use...We really get no automatic notification that a patient has died. They die elsewhere, and the state finds out because there's a death certificate somewhere, but our HIM [health information management] department is sort of stuck almost to the point of reading obituaries trying to figure out what patients to mark...It's really hard, from our perspective to reach out to a family with an appointment reminder about a patient who died, and it's just, not only is it horrible customer service and patient experience or family experience to do that...I think more connection points with actually the state for some of this basic stuff like birth records and death records and marriage certificates where names are changing and that information's sitting there, but it seems to be behind this kind of either bureaucratic or policy firewall." [Hospital representative #15]
Exchanging data across the continuum of care	"In our long-term care systems, it would be nice for us to be able to exchange information about those patients, especially with medications, make sure their medications are set up and they know everything that the patient's on. One thing, too, is provide our providers with information, a little bit more timely from the long-term care facilities to keep them from being readmitted or admitted to the hospital. So that's something we'd like to be able to do. Another thing would be the merging of medical records for both mental health and their regular healthcare." [Primary care provider #6]
Reliance on regional HIEs[a] for data exchange	"I see a lot of potential for us to be able to really be part of that health information exchange network, and use our EHR [electronic health record] system to do a lot of that in the background, as opposed to currently what we are doing is we have access to [regional HIE] portal to get the community health record, the information, but that takes staff time. You know, a lot of training, and so it's not really fully integrated into our EHR [electronic health record], and it's not fully integrated into our clinical practice processes." [Behavioral health provider #17]
Opportunities	
Using new population health software to improve interoperability	"That's the beauty of it. So, I...mentioned computer vision[b], so what that's compared to if you wanted to do that in the past, you would have had to have a pretty labor-intensive interface between the platform and each individual practice's EHR [electronic health record], right. And that's a huge level of effort that most people can't really get to, and that's why the computer vision piece of that really makes sense. You don't have to have that, well, it's still fancy, it's fancy in a different way, you don't need a fancy interface, you're using the computer vision[a] to match the patients." [Hospital representative #21]
Integrating care coordination programs to improve social needs referrals	"Back to the social determinants...the opportunity to connect to external things like Aunt Bertha or NowPow or Healthify, one of those products that helps kind of do closed loop referral, and whether we'll do that within Epic...I think those...are helping us kind of reach our goals around reducing health disparities." [Primary care provider #14]
Using FHIR[c]-based applications to advance interoperability	"We are just implementing our FHIR [Fast Healthcare Interoperability Resources] layer right now. What I will tell you is while FHIR [Fast Healthcare Interoperability Resources] is definitely a direction of the future, it is not broadly deployed in the marketplace and not broadly deployed in the workflow or business applications to great extent. But it is, definitely will be an important factor as we move into the future. But it also will not be the silver bullet that everybody's hoping it was going to be." [HIE representative #8]

[a]HIE: health information exchange.

[b]Computer vision is a field of technology that enables devices such as smart cameras to acquire, process, analyze, and interpret text, images, and videos.

[c]FHIR: Fast Healthcare Interoperability Resources.

Implementation Challenges

For hospitals, a primary concern focused on the growing number and proliferation of application types (eg, EHR, personal health records, HIE, and population health platforms) that contain health information and can be used in clinical encounters. These applications were viewed as difficult and costly to maintain. Similarly, these applications do not use a common data structure and storage format, resulting in too many places for data to be located and for clinicians to search for useful information. The lack of interoperability necessitates that providers leave the EHR to access other applications. As one hospital representative stated as follows:

> *...It's one of those things where now you've got to subscribe to it [eg, a regional HIE], and it's another place [eg, application] for you to go to look for more information [eg, patient clinical data]...There's too many places for data to land and get sent. People just stop looking.* [Hospital representative #10]

HIE leadership echoed this perspective, adding that a lack of interoperability was an impediment to creating a complete medical record. PCPs noted similar issues related to the number and type of applications needed to overcome gaps in interoperability, such as population health tools (eg, Innovaccer). In contrast, for BHPs, interoperability across applications was not discussed; instead, remarks focused on off-the-shelf EHR systems being misaligned to the specific requirements for BHPs, such as lacking additional protections for substance abuse data.

Interoperability Capabilities

Hospitals described advanced interoperability functionality and attributed their advanced capabilities to their EHR vendor rather than the regional HIEs. The regional HIEs were helpful for exchanging continuity of care documents [27] but did not facilitate the integration of information across EHRs. Typically, hospitals only expressed limitations with interoperability functionality as being a function of data recipients' capacities. For instance, a major concern for hospitals was connecting and being able to integrate EHR data with state-hosted information systems such as the state immunization registry, vital statistics, death certificates, or birth records.

Hospitals and PCPs both noted limitations of exchanging data across the continuum of care, such as with long-term care and BHPs. PCPs additionally noted interoperability challenges with hospitals in their own health systems even when they all used the same EHR vendor, as some vendors are not capable of data exchange in different instances of the same EHR.

BHPs described much more basic data exchange capabilities relative to hospitals and PCPs. For instance, they mentioned that their data exchange is primarily focused on billing. In contrast to hospitals, BHPs mentioned a greater reliance on the regional HIEs for access to health records from other providers and event notifications. The regional HIEs echoed this relationship and discussed how BHPs lag in their interoperability capabilities.

Opportunities

Hospitals described opportunities related to new population health platforms that may be able to improve interoperability without the costs associated with interfacing with different EHR systems or HIEs. Both hospitals and PCPs felt that increasing analytic rigor and predictive modeling with artificial intelligence and machine learning would support their population health efforts. API-based data exchange was generally identified as an opportunity to promote interoperability. Hospitals mentioned that the use of applications with this technology will benefit remote patient monitoring and chronic care management. API-based data sharing was also identified as an opportunity to improve interoperability in behavioral health despite the focus on more basic technological opportunities, such as increasing EHR functionality.

The regional HIEs also shared these perspectives on API-based data sharing and were particularly attentive to FHIR APIs for the development and implementation of EHR-integrated applications to advance interoperability. FHIR-based applications were also expected to improve patient access to their medical records. Although most hospitals and PCPs offered patient access to their medical records through patient portals at the time of our study, these were not unified across providers. FHIR-based applications could potentially enable broader patient access through a unified patient portal that collects information from disparate providers. However, the regional HIEs tempered this enthusiasm, recognizing that FHIR-based applications are not currently broadly deployed in existing technology builds.

Hospitals, PCPs, and HIEs also discussed using care coordination programs (eg, Aunt Bertha) that can track referrals to social service agencies that address social determinants of health (SDoH). These providers described collecting SDoH data in discrete data fields but also noted that the lack of standardization of SDoH data fields remains problematic and results in questionable data quality and limited ability to combine data across sources. Efforts to develop SDoH data standards, such as the Gravity Project [28], were raised as opportunities to improve the interoperability of SDoH data.

Organization Domain

Overview

We identified two themes within the organizational domain that affected interoperability: (1) facilitators and (2) strategic alignment (Table 3).

Table 3. Themes and primary subthemes in the organization domain.

Themes and primary subthemes	Representative quote
Facilitators	
Relationship with EHR[a] vendor	"They [EHR vendor] are as interested in making sure that interoperability happens as what we are and so when we start to look at different interfaces that we need to have built, whether it's to another vendor, like Cerner or eClinicalWorks, then Epic builds that interface, so that it makes it easier on both ends, to make that connection." [Hospital representative #3]
Data standards adoption	"I think standards, the general and the meaningful use did quite a bit of pushing this sort of embracing of standards around things like nomenclatures, terminologies that allow for transmission of information. Prior to this we're pretty much stuck with HL7 [Health Level 7] and custom specifications, but now with just CPT [current procedural technology] or ICD [International Classification of Disease] but, between SNOMED [Systematized Nomenclature of Medicine] and LOINC [Logical Observation Identifiers, Names and Codes] and RxNorm and CVX [vaccine administered] codes for immunizations and it's gotten a lot better." [Hospital representative #15]
Senior leadership support	"The biggest thing [to support data gathering and integration] is the addition of scribes. Adding on that expense of additional manpower to do that data entry for the providers to get them to where they're comfortable with...what is pertinent to that visit. It may be an ER [emergency room] visit, is it a recent CT [computerized tomography scan], so that way staff isn't trying to print the last X-amount of things and really all they wanted was one thing, so trying to streamline that to get the physicians the information they need, but having a team around them, to help them put the information back in and alleviate that work from them." [Primary care provider #16]
Strategic alignment	
Payment program participation impacts technology purchasing decisions	"We are in CPC [Comprehensive Primary Care]-Ohio and CPC+ [Comprehensive Primary Care Plus]. We're also doing Primary Care First and I have a number of value-based commercial contracts that we deal with as well. We are not an ACO [accountable care organization]. Our new software with our population health software that's been added to our regular EMR [electronic medical record] should help greatly with that and that's the reason we did it is because we're getting into more value-based contracts. I think that's something that will improve our outcomes for our patients and improve our financial return as well." [Primary care provider #6]
Using interoperability to develop cross-sector alignment and stakeholder consensus	"I'd love to see us in the state of Ohio come together at kind of a developer's conference or something...How can we come together and figure out how to make this work better for Ohio? And I know that sounds really altruistic because everyone's trying to run a business and all that, but it just seems like there's so much overlap and you think—I'll use the example with us: I'm sitting on a mountain of data. Right, and so it just drives me bonkers to hear of a small mom-and-pop startup software company, who has to go out and buy a big giant data warehouse, you know, big giant SQL server and pay licenses and then they contact all the hospitals in the doctor's office and say, give me all your data, right. And we're just duplicating these silos. Not too long ago, I was giving a presentation, I said, how many of your hospital systems have a population health strategy? And, of course, 100% of them raised their hand, right. And I said so you've invested millions into giant data warehouses to support population health, you know. Right? And they all go, yeah. Like, well so did all the HIEs[b] [health information exchanges], so did the state of Ohio, you know. ODH [Ohio Department of Health] is trying to, like we're all we're all spending—Microsoft and Oracle and all those guys are making money hand over fist. Because we can't get ourselves organized." [HIE representative #4]
Different perspectives on the value of interoperability	"They didn't see how interoperability would help them take care of their patients any better. And our team even said, 'Well, you can get lab results like instantaneously.' 'Yeah, you know, but I get them a day late. It's fine.' ...I think that's the other challenge is really, is that one example, or is that some X percent of providers in the state of Ohio who don't see value in that interoperability." [HIE representative #4]

[a]EHR: electronic health record.

[b]HIE: health information exchange.

Facilitators

A consistent subtheme related to the importance of relationships with the EHR vendor to support interoperability emerged. Hospitals placed considerable value on their relationship with Epic and perceived the high concentration of state-wide Epic institutions and the integration between HIEs and Epic as a benefit. Hospitals also mentioned the push for data standards through meaningful use as facilitating interoperability.

PCPs described the important role of senior leadership, who can designate sufficient human resources for tasks that typically increase clinical workload, such as data gathering before appointments. These staff resources can help physicians access and use information from other sources.

Strategic Alignment

A critical driver for interoperability among hospitals was their participation in value-based payment programs and population health initiatives. To facilitate the data exchange and integration

for care coordination, billing, and reporting required to support these programs, hospitals reported purchasing EHRs for network partners that lacked these advanced capabilities, thus promoting interoperability among their clinical partners. Likewise, hospitals preferred to develop their own in-house population health analytics platforms rather than outsource this function to HIEs.

PCPs also felt that interoperability is central to achieving strategic goals such as population health management, which is integral to participation in alternative payment models (ie, accountable care organizations, CPC-Ohio, CPC-plus, and Primary Care First). They noted the benefits of event notifications made possible by admission, discharge, and transfer feeds that allow PCPs to be notified when their patients have received care in other settings and follow-up accordingly with them to meet their needs.

Conversely, BHPs indicated that they were not participating in value-based purchasing programs to the same degree. However, similar to hospitals and PCPs, their technology investment decisions were guided by their participation in payment programs with sponsors (eg, Health Resources and Services Administration and Substance Abuse and Mental Health Services Administration) that require specific reports, although they may not necessarily aid in interoperability.

The regional HIEs viewed the development of their exchange networks as an opportunity to advocate for cross-sector alignment, particularly as it pertains to streamlining duplicate efforts toward population health management. They remarked that regional HIEs are in a unique position to negotiate partnerships that address the concerns of different stakeholders. Finally, the regional HIEs noted challenges related to some organizations, particularly BHPs, not viewing interoperability as valuable to their organization or aligning with their strategy.

Environment Domain

Within the environmental domain, a policy theme was identified that focused on how policy may hinder or facilitate interoperability (Table 4).

Hospitals felt that there was room for a greater policy push for managed care plans to initiate value-based payment contracts and distribute incentives to providers. Hospitals also felt the costs of interoperability, such as establishing admission, discharge, and transfer feeds or registry reporting, fall predominantly on hospitals, yet there are no corresponding changes to reimbursement. Similarly, BHPs also desired additional funding to support their adoption of advanced EHR systems.

All providers, including the regional HIEs, also noted the considerable impact of 42 Code of Federal Regulations (CFR) Part 2 (ie, Substance Abuse Confidentiality Regulation) as limiting interoperability and connectivity with BHPs. Despite the general recognition of the policy's good intentions, it was viewed as negatively impacting clinical care. Some key informants felt that this policy was incongruent with new information blocking rules as part of the 21CCA. The regional HIEs also saw consequences of this policy with respect to responding to public health emergencies such as the opioid epidemic. BHPs, PCPs, and the regional HIEs all advocated for aligning 42 CFR Part 2 with the HIPAA to clarify what is protected and not shareable versus what can be shared for continuity of care.

Table 4. Themes and primary subthemes in the environment domain.

Policy	Representative quote
Push insurance plans to participate in value-based payment programs	"The way to actually advance Triple Aim[a], right, is, is to have these accountable care organizations and really strongly incent the provider organizations to get on board with them. Well, if it's like pulling teeth to get the Medicaid [insurance plans] to work with us on that, then that's limited. And so I haven't seen a lot of folks...in Columbus really encouraging, like, what are they doing to really encourage that those accountable care incentives get down to the provider level. So there's the CPC+ [Comprehensive Primary Care Plus] program, which is good to a point, because that gives primary care providers some funds upfront to theoretically invest in all this stuff. We're experiencing now is [Medicaid insurance plans] came on board for [ACO[b] Network], but none of the other providers are, and they're coming up with excuses why they don't want to do it. It's like, how come someone in Columbus isn't telling them to get on board with a [ACO Network] program." [Hospital representative #21]
Increase funding for adoption of interoperable EHR[c]	"Most people don't realize going into an implementation is how much it will cost you...more of those incentives, I think, would be beneficial even if it's for individuals moving to a better record that will allow them to do more of the communication between systems and all that. Our current system will never have that capability." [Behavioral health provider #18]
Substance abuse confidentiality requirements (42 CFR[d] Part 2) limit interoperability with behavioral health providers	"When you go back to advocacy, the biggest thing we need to do around that in my world again is to align the Part 2 information with more the HIPAA[e] [Health Insurance Portability and Accountability Act] guidelines, so we can make sure that that information gets out to the primary care providers, gets out to those entities that are trying to support these people. I think that's one of the biggest roadblocks to taking the next step and helping with the opioid crisis, because again, I think Ohio has done some great work here, but our hands are kind of tied right now, because of the federal rules. And so advocating through the state up to the feds to better align those rules so information can be shared, is probably pretty key." [HIE[f] representative #8]
Align 42 CFR Part 2 with HIPAA to clarify what is protected data and what can be shared	"We would love to see movement on the alignment with 42 CFR [Code of Federal Regulations] and HIPAA [Health Insurance Portability and Accountability Act], which would allow sharing without having to always parse out." [Behavioral health provider #7]
Information blocking continues to exist	"I think the feds really need to step up some of their efforts a little bit, and I've been pretty vocal with folks at ONC [Office of the National Coordinator] in DC is data sharing is one part of that information blocking conversation. We've got to got to do a better job really supporting that and pushing on EMR [electronic medical record] vendors to send all the data. It makes it very, very challenging to match up data if somebody says, I can't send you addresses. Well then, that data is almost worthless when you're talking about trying to track that by zip code and say, hey public health, you're having an outbreak in zip code 12345. So, if you don't get addresses from some of the EMRs [electronic medical records], you know, there's not enough expectation I don't think even in the HL7 [Health Level 7] standards and some of the others." [HIE representative #4]
TEFCA[g] is helpful to establish standards, but could be expanded to promote adoption of FHIR[h]	"And to my point, the national effort, right now, that is looking at expanding the national scope of exchange, it came out of the Cures Act, called TEFCA, the Trusted Exchange and Common Agreement Framework, the standards that they have just pushed out do not include FHIR [Fast Healthcare Interoperability Resources] as a standard...And again, don't get me wrong, it is going to be the way of the future, and you have to be able to integrate that into your technology stack, but we've got a ways to go." [HIE representative #8]

[a]Triple Aim refers to an approach to optimizing health system performance based on improving population health, enhancing the care experience, and reducing costs [29].

[b]ACO: accountable care organization.

[c]EHR: electronic health record.

[d]CFR: Code of Federal Regulations.

[e]HIPAA: Health Insurance Portability and Accountability Act.

[f]HIE: health information exchange.

[g]TEFCA: Trusted Exchange Framework Common Agreement.

[h]FHIR: Fast Healthcare Interoperability Resources.

The regional HIEs noted some specific concerns, given their experience in the COVID-19 response. First, they described a need for greater enforcement of information-blocking rules. The implications of information blocking were particularly notable during the pandemic when withheld information about addresses prevented HIE from tracking COVID-19 at the local level. Second, the regional HIEs felt that HIPAA guidance on reporting on geographic areas with less than 20,000 patients is both vague and confusing, limiting population health efforts.

Finally, the regional HIEs also mentioned that TEFCA helps to establish standards but did not include any information about FHIR-based standards. They noted that this update would be helpful to promote adoption of FHIR-based applications [30].

Discussion

Principal Findings

The near-ubiquitous adoption of certified EHRs over the past decade has resulted in the capture of vast amounts of data across the care continuum. Recent policy efforts such as the 21CCA aim to promote vendor-agnostic integration of external data into the EHRs of all provider types and address information blocking to mitigate challenges. However, questions persist about how well these efforts assist health care organizations in achieving interoperability. Our study examined perspectives from a variety of provider types in Ohio on the state of interoperability. We identified important barriers and facilitators to interoperability among hospitals, PCPs, BHPs, and regional HIEs.

Our findings related to implementation issues within the technology domain suggest that the proliferation of applications that address various use cases promotes capture of rich data.

However, from an end user perspective, this approach can create inefficiencies because of excess information. User interfaces that do not embed multiple discrete applications within the EHR may ultimately create a fragmented medical record, which makes it harder to find relevant information. These types of information silos of patient data can potentially jeopardize patient safety, care quality, and organizational efficiency [31,32]. Fragmented or siloed information may also contribute to provider burnout [33]. Thus, our findings highlight a need for user-centric approaches in technology design and implementation to translate increased information access to use.

Providers did note that technological advances such as computer vision-based population health software (eg, Innovaccer) can help overcome these barriers by pulling data directly into the EHR without requiring back-end integration. Likewise, the growing support for API-based data sharing can further support interoperable information exchange between dissimilar EHR vendors. APIs can extend EHR capabilities [34], although this potential remains unrealized to date. Recent initiatives such as TEFCA that mandate sharing standardized sets of data and promote the use of FHIR APIs are expected to facilitate interoperability in the coming years [30]. Moreover, the API approach has the potential to facilitate population health analytics using machine learning techniques [35].

Within the technology domain, our findings reiterated a gap between hospitals and PCPs at one end of the spectrum and BHPs at the other with respect to their interoperability capabilities. This gap may be a result of most BHPs being ineligible for federal incentive programs [18]. This omission likely not only disincentivizes the adoption of advanced EHRs capable of interoperability among BHPs but also discourages investment and development by EHR vendors of tools designed to meet the needs of BHPs [36].

In addition, across the technology, organization, and environment domains and across provider types, a notable issue that emerged from the key informant interviews was the limitation on interoperability imposed by 42 CFR Part 2, a rule that restricts sharing substance use and behavioral health data. Our findings suggest that despite its intentions, 42 CFR Part 2 effectively operates to prevent, or severely limit, BHPs from participating in exchange. In addition to potential medical errors, this restriction may result in missing data from analytic data sets used by providers, insurers, or researchers [37]. Further, providers expressed concern that this rule may no longer be in alignment with 21CCA information-blocking rules. Moving forward, modifications to 42 CFR Part 2 may be necessary to support further interoperability. Consideration of end user needs and incorporating perspectives of BHPs are essential in any policy changes to carefully consider the need for privacy related to substance use and behavioral health data balanced against the benefits of interoperability [38]. To this effect, the United States Department of Health and Human Services issued a Notice of Proposed Rulemaking in November 2022 [39]. The Notice of Proposed Rulemaking proposes to permit the use and disclosure of Part 2 records based on a single prior signed consent, to expand prohibitions on the use and disclosure of Part 2 records in legal proceedings and to expand patient rights that align with the HIPAA Privacy Rule. The importance of this issue has recently become more visible because of concerns around the protections of reproductive health information following increased abortion restrictions because of the overturning of the *Roe versus Wade* supreme court case [40,41]. The revisions to Part 2, if enacted, would not only help BHPs engage in interoperability but would also provide greater protection for how sensitive health data can be used in legal proceedings.

Interestingly, hospitals attributed their interoperability capabilities more to their relationship with EHR vendors as opposed to the regional HIEs. This finding is likely a consequence of the strong foothold of Epic in Ohio. From an operational and governance standpoint, providers may face fewer barriers to participating in Epic's Care Everywhere vendor-mediated HIE network. Conversely, participation with a regional HIE may require further effort to establish data exchange policies that require buy-in from multiple stakeholders. Vendor-mediated HIE may address barriers to use, such as the need to leave the EHR to access information. Nonetheless, these vendor-mediated HIEs may create a divide in HIE participation between providers with different vendors [42]. Indeed, as the HITECH funding period drew to a close, the number of state and regional HIEs declined, partly because of the mergers of regional HIEs, funding challenges, and competition from vendor-mediated HIEs [43]. Vendor-mediated and regional HIEs are not necessarily mutually exclusive, although regional HIEs may be more inclusive of a range of provider types. Policies targeted at supporting regional HIEs may be necessary to counteract the market forces of vendor-mediated HIEs and keep interoperability obtainable for nonhospital or hospital-affiliated practices.

Our findings related to strategic alignment may offer useful policy recommendations; provider participation in value-based payment programs plays a critical role in how providers are considering (or not considering) investments in interoperability. To the extent that most of a provider's patients are beneficiaries of these programs, providers may expand the breadth of their interoperability functionality, such as participation in regional HIEs or use of population health platforms to meet the needs

of that particular patient population [44]. Key informants also described the use of social needs referral platforms; however, developing exchange with non-HIPAA–covered social service or community-based organizations can be challenging without properly aligned incentives [45]. In addition to clarifying HIPAA rules around exchange with noncovered entities, the expansion of value-based payment programs may leverage HIT investment. Further, regional HIEs may be well positioned to advocate for cross-sector strategic alignment.

The providers in our study reported capacity-related challenges in interoperability with public health agencies. Chronically underfunded public health systems impeded efficient and timely electronic information exchange during the COVID-19 pandemic [46]. In response, the Centers for Disease Control and Prevention launched the Data Modernization Initiative, resulting in changes to core data sources and facilitating access to electronic case reports and the COVID electronic laboratory reporting that makes test results available. Other initiatives such as TEFCA, through their emphasis on interoperability, are also expected to mitigate challenges, particularly from differing vocabulary standards. Other barriers to interoperability in public health arise from the complex legal and regulatory environment [47]. Even though 21CCA established a legal framework to address information blocking [48], the HIEs participating in our study reported information blocking that prevented tracking cases of COVID-19. Moving forward, it will be critical to monitor the impact of the Data Modernization Initiative and TEFCA on the interoperability of public health data.

Limitations

We purposely sampled from multiple stakeholders to gain a representative perspective on interoperability. The design focused on breadth across providers rather than depth within a specific provider type or within a single health care organization. Similarly, the study only included Ohio stakeholders, which minimizes variation in the policy environment and may limit generalizability.

The sampling approach was designed to include individuals with decision-making authority with respect to HIT. These perspectives may differ from those of other end users. Finally, owing to time and resource constraints, the interview guide may not have probed all issues relevant to interoperability but was purposefully open-ended to allow participants to discuss topics they deemed important.

Conclusions

Our findings suggest that despite the ubiquity of data and applications, seamless interoperability into a comprehensive medical record, both within and across providers, remains out of reach. Technological solutions offer promise to overcome these challenges. Likewise, the expansion of value-based payment programs can further incentivize interoperability. Although policy initiatives to expand interoperability existed, they were often misaligned to operational needs and may not be sufficient to overcome market forces. A policy focus toward embracing user-centric design to incorporate end user experience into HIT development may overcome barriers associated with achieving interoperability.

Acknowledgments

This project was supported by a contract from the Ohio Department of Medicaid. The content is solely the responsibility of the authors and does not necessarily represent the official views of the sponsor. The authors would like to thank Lauren Phelps for her assistance with project management.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Interview guides.
[DOCX File , 41 KB - medinform_v11i1e43848_app1.docx]

References

1. Akhlaq A, Sheikh A, Pagliari C. Defining health information exchange: scoping review of published definitions. J Innov Health Inform 2017 Jan 25;23(4):838 [FREE Full text] [doi: 10.14236/jhi.v23i4.838] [Medline: 28346130]
2. Fridsma D. Interoperability vs health information exchange: setting the record straight. Health IT Buzz. 2013 Jan 9. URL: https://tinyurl.com/jefef6mm [accessed 2022-04-29]
3. Bates DW, Samal L. Interoperability: what is it, how can we make it work for clinicians, and how should we measure it in the future? Health Serv Res 2018 Oct;53(5):3270-3277 [FREE Full text] [doi: 10.1111/1475-6773.12852] [Medline: 29527678]
4. Buntin MB, Jain SH, Blumenthal D. Health information technology: laying the infrastructure for national health reform. Health Aff (Millwood) 2010 Jun;29(6):1214-1219. [doi: 10.1377/hlthaff.2010.0503] [Medline: 20530358]
5. Williams C, Mostashari F, Mertz K, Hogin E, Atwal P. From the office of the National Coordinator: the strategy for advancing the exchange of health information. Health Aff (Millwood) 2012 Mar;31(3):527-536. [doi: 10.1377/hlthaff.2011.1314] [Medline: 22392663]

6. Jackson CT, Trygstad TK, DeWalt DA, DuBard CA. Transitional care cut hospital readmissions for North Carolina medicaid patients with complex chronic conditions. Health Aff (Millwood) 2013 Aug;32(8):1407-1415 [FREE Full text] [doi: 10.1377/hlthaff.2013.0047] [Medline: 23918485]
7. Peikes D, Dale S, Ghosh A, Taylor EF, Swankoski K, O'Malley AS, et al. The comprehensive primary care initiative: effects on spending, quality, patients, and physicians. Health Aff (Millwood) 2018 Jun;37(6):890-899 [FREE Full text] [doi: 10.1377/hlthaff.2017.1678] [Medline: 29791190]
8. Singh P, Orzol S, Peikes D, Oh EG, Dale S. Participation in the comprehensive primary care plus initiative. Ann Fam Med 2020 Jul;18(4):309-317 [FREE Full text] [doi: 10.1370/afm.2544] [Medline: 32661031]
9. Primary care first model options. Centers for Medicare & Medicaid Services. URL: https://innovation.cms.gov/innovation-models/primary-care-first-model-options [accessed 2022-12-16]
10. Henry J, Pylypchuk Y, Searcy T, Patel V. Adoption of electronic health record systems among U.S. non-federal acute care hospitals: 2008-2015. The Office of the National Coordinator for Health Information Technology. 2016 May. URL: https://www.healthit.gov/sites/default/files/briefs/2015_hospital_adoption_db_v17.pdf [accessed 2022-04-15]
11. Office-based physician electronic health record adoption. Office of the National Coordinator for Health Information Technology. 2019. URL: https://www.healthit.gov/data/quickstats/office-based-physician-electronic-health-record-adoption [accessed 2022-04-20]
12. Rahurkar S, Vest JR, Finnell JT, Dixon BE. Trends in user-initiated health information exchange in the inpatient, outpatient, and emergency settings. J Am Med Inform Assoc 2021 Mar 01;28(3):622-627 [FREE Full text] [doi: 10.1093/jamia/ocaa226] [Medline: 33067617]
13. Payne TH, Corley S, Cullen TA, Gandhi TK, Harrington L, Kuperman GJ, et al. Report of the AMIA EHR-2020 task force on the status and future direction of EHRs. J Am Med Inform Assoc 2015 Sep;22(5):1102-1110 [FREE Full text] [doi: 10.1093/jamia/ocv066] [Medline: 26024883]
14. Institute of Medicine, Board on Health Care Services, Committee on Patient Safety and Health Information Technology. Health IT and Patient Safety: Building Safer Systems for Better Care. Washington, DC, USA: National Academies Press; 2011.
15. Roman LC, Ancker JS, Johnson SB, Senathirajah Y. Navigation in the electronic health record: a review of the safety and usability literature. J Biomed Inform 2017 Mar;67:69-79 [FREE Full text] [doi: 10.1016/j.jbi.2017.01.005] [Medline: 28088527]
16. Holmgren AJ, Everson J, Adler-Milstein J. Association of hospital interoperable data sharing with alternative payment model participation. JAMA Health Forum 2022 Feb;3(2):e215199 [FREE Full text] [doi: 10.1001/jamahealthforum.2021.5199] [Medline: 35977275]
17. Cross DA, Stevens MA, Spivack SB, Murray GF, Rodriguez HP, Lewis VA. Survey of information exchange and advanced use of other health information technology in primary care settings: capabilities in and outside of the safety net. Med Care 2022 Feb 01;60(2):140-148. [doi: 10.1097/MLR.0000000000001673] [Medline: 35030563]
18. Walker D, Mora A, Demosthenidy MM, Menachemi N, Diana ML. Meaningful use of EHRs among hospitals ineligible for incentives lags behind that of other hospitals, 2009-13. Health Aff (Millwood) 2016 Mar;35(3):495-501 [FREE Full text] [doi: 10.1377/hlthaff.2015.0924] [Medline: 26953305]
19. Shields MC, Horgan CM, Ritter GA, Busch AB. Use of electronic health information technology in a national sample of hospitals that provide specialty substance use care. Psychiatr Serv 2021 Dec 01;72(12):1370-1376 [FREE Full text] [doi: 10.1176/appi.ps.202000816] [Medline: 33853380]
20. Pronovost P, Johns MM, Palmer S. Procuring Interoperability: Achieving High-Quality, Connected, and Person-Centered Care. Washington, DC, USA: National Academy of Medicine; 2018.
21. Adler-Milstein J, Garg A, Zhao W, Patel V. A survey of health information exchange organizations in advance of a nationwide connectivity framework. Health Aff (Millwood) 2021 May;40(5):736-744 [FREE Full text] [doi: 10.1377/hlthaff.2020.01497] [Medline: 33939510]
22. HiTech environmental scan: final report. Government Resource Center, Ohio Colleges of Medicine. 2022 Jan. URL: http://grc.osu.edu/search/node?keys=HiTech+Environmental+Scan%3A+Final+Report [accessed 2022-10-30]
23. Tornatzky LG, Fleischer M, Chakrabarti AK. The Processes of Technological Innovation. Lexington, MA, USA: Lexington Books; 1990.
24. Walker DM, Yeager VA, Lawrence J, McAlearney AS. Identifying opportunities to strengthen the public health informatics infrastructure: exploring hospitals' challenges with data exchange. Milbank Q 2021 Jun;99(2):393-425 [FREE Full text] [doi: 10.1111/1468-0009.12511] [Medline: 33783863]
25. Hsieh HF, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005 Nov;15(9):1277-1288 [FREE Full text] [doi: 10.1177/1049732305276687] [Medline: 16204405]
26. White DE, Oelke ND, Friesen S. Management of a large qualitative data set: establishing trustworthiness of the data. Int J Qual Methods 2012 Jul 01;11(3):244-258 [FREE Full text] [doi: 10.1177/160940691201100305]
27. Ferranti JM, Musser RC, Kawamoto K, Hammond WE. The clinical document architecture and the continuity of care record: a critical analysis. J Am Med Inform Assoc 2006;13(3):245-252 [FREE Full text] [doi: 10.1197/jamia.M1963] [Medline: 16501180]

28. Lousberg C, Behal S. The Gravity Project. Confluence. 2022. URL: https://confluence.hl7.org/display/GRAV/The+Gravity+Project [accessed 2022-04-22]
29. Berwick DM, Nolan TW, Whittington J. The triple aim: care, health, and cost. Health Aff (Millwood) 2008;27(3):759-769 [FREE Full text] [doi: 10.1377/hlthaff.27.3.759] [Medline: 18474969]
30. FHIR® Roadmap for TEFCA Exchange: Version 1. Office of the National Coordinator for Health Information Technology. 2022. URL: https://rce.sequoiaproject.org/wp-content/uploads/2022/01/FHIR-Roadmap-v1.0_updated.pdf [accessed 2022-04-22]
31. Payne T, Fellner J, Dugowson C, Liebovitz D, Fletcher G. Use of more than one electronic medical record system within a single health care organization. Appl Clin Inform 2012 Dec 12;3(4):462-474 [FREE Full text] [doi: 10.4338/ACI-2012-10-RA-0040] [Medline: 23646091]
32. Ahmed Z, Jani Y, Franklin BD. Qualitative study exploring the phenomenon of multiple electronic prescribing systems within single hospital organisations. BMC Health Serv Res 2018 Dec 14;18(1):969 [FREE Full text] [doi: 10.1186/s12913-018-3750-1] [Medline: 30547779]
33. Kroth PJ, Morioka-Douglas N, Veres S, Pollock K, Babbott S, Poplau S, et al. The electronic elephant in the room: physicians and the electronic health record. JAMIA Open 2018 Jul;1(1):49-56 [FREE Full text] [doi: 10.1093/jamiaopen/ooy016] [Medline: 31093606]
34. Gordon WJ, Rudin RS. Why APIs? Anticipated value, barriers, and opportunities for standards-based application programming interfaces in healthcare: perspectives of US thought leaders. JAMIA Open 2022 Jul;5(2):ooac023 [FREE Full text] [doi: 10.1093/jamiaopen/ooac023] [Medline: 35474716]
35. Huber L, Honeder T, Hackl WO. FHIR analytics - pragmatic review of recent studies. Stud Health Technol Inform 2020 Jun 23;271:110-112 [FREE Full text] [doi: 10.3233/SHTI200083] [Medline: 32578550]
36. Ranallo PA, Kilbourne AM, Whatley AS, Pincus HA. Behavioral health information technology: from chaos to clarity. Health Aff (Millwood) 2016 Jun 01;35(6):1106-1113 [FREE Full text] [doi: 10.1377/hlthaff.2016.0013] [Medline: 27269029]
37. Madden JM, Lakoma MD, Rusinak D, Lu CY, Soumerai SB. Missing clinical and behavioral health data in a large electronic health record (EHR) system. J Am Med Inform Assoc 2016 Nov;23(6):1143-1149 [FREE Full text] [doi: 10.1093/jamia/ocw021] [Medline: 27079506]
38. Frakt AB, Bagley N. Protection or harm? Suppressing substance-use data. N Engl J Med 2015 May 14;372(20):1879-1881 [FREE Full text] [doi: 10.1056/NEJMp1501362] [Medline: 25875196]
39. HIPAA and Part 2: 42 CFR Part 2 Rulemaking. Office of Civil Rights, Department of Health and Human Services. 2022. URL: https://www.hhs.gov/hipaa/for-professionals/regulatory-initiatives/hipaa-part-2/index.html [accessed 2022-12-16]
40. Walker DM, Hoffman S, Adler-Milstein J. Interoperability in a post-roe era: sustaining progress while protecting reproductive health information. JAMA 2022 Nov 01;328(17):1703-1704 [FREE Full text] [doi: 10.1001/jama.2022.17204] [Medline: 36318125]
41. Spector-Bagdady K, Mello MM. Protecting the privacy of reproductive health information after the fall of Roe v Wade. JAMA Health Forum 2022 Jun 03;3(6):e222656 [FREE Full text] [doi: 10.1001/jamahealthforum.2022.2656] [Medline: 36219024]
42. Everson J, Barker W, Patel V. Electronic health record developer market segmentation contributes to divide in physician interoperable exchange. J Am Med Inform Assoc 2022 Jun 14;29(7):1200-1207 [FREE Full text] [doi: 10.1093/jamia/ocac056] [Medline: 35442438]
43. Adler-Milstein J, Lin SC, Jha AK. The number of health information exchange efforts is declining, leaving the viability of broad clinical data exchange uncertain. Health Aff (Millwood) 2016 Jul 01;35(7):1278-1285 [FREE Full text] [doi: 10.1377/hlthaff.2015.1439] [Medline: 27385245]
44. Apathy NC, Holmgren AJ, Werner RM. Growth in health information exchange with ACO market penetration. Am J Manag Care 2022 Jan 01;28(1):e7-13 [FREE Full text] [doi: 10.37765/ajmc.2022.88815] [Medline: 35049261]
45. Walker DM, Hefner JL, DePuccio MJ, Garner JA, Headings A, Joseph JJ, et al. Approaches for overcoming barriers to cross-sector data sharing. Am J Manag Care 2022 Jan;28(1):11-16 [FREE Full text] [doi: 10.37765/ajmc.2022.88811] [Medline: 35049256]
46. Dixon BE, Caine VA, Halverson PK. Deficient response to COVID-19 makes the case for evolving the public health system. Am J Prev Med 2020 Dec;59(6):887-891 [FREE Full text] [doi: 10.1016/j.amepre.2020.07.024] [Medline: 32978011]
47. Dixon BE, Rahurkar S, Apathy NC. Interoperability and health information exchange for public health. In: Magnuson JA, Dixon BE, editors. Public Health Informatics and Information Systems. Cham, Switzerland: Springer; 2020.
48. Black JR, Hulkower RL, Ramanathan T. Health information blocking: responses under the 21st Century Cures Act. Public Health Rep 2018 Sep;133(5):610-613 [FREE Full text] [doi: 10.1177/0033354918791544] [Medline: 30134128]

Abbreviations

API: application programming interface
BHP: behavioral health provider
CFR: Code of Federal Regulations

CPC: Comprehensive Primary Care
EHR: electronic health record
FHIR: Fast Healthcare Interoperability Resources
HIE: health information exchange
HIPAA: Health Insurance Portability and Accountability Act
HIT: health information technology
HITECH: Health Information Technology for Economic and Clinical Health
MPIP: Medicaid Provider Incentive Program
PCP: primary care provider
SDoH: social determinants of health
TEFCA: Trusted Exchange Framework Common Agreement
21CCA: 21st Century Cures Act

Edited by A Benis; submitted 27.10.22; peer-reviewed by S Tian, O Petrovskaya; comments to author 26.11.22; revised version received 11.01.23; accepted 19.01.23; published 24.02.23.

Please cite as:
Walker DM, Tarver WL, Jonnalagadda P, Ranbom L, Ford EW, Rahurkar S
Perspectives on Challenges and Opportunities for Interoperability: Findings From Key Informant Interviews With Stakeholders in Ohio
JMIR Med Inform 2023;11:e43848
URL: https://medinform.jmir.org/2023/1/e43848
doi:10.2196/43848
PMID:36826979

Original Paper

The Effect of Implementation of Guideline Order Bundles Into a General Admission Order Set on Clinical Practice Guideline Adoption: Quasi-Experimental Study

Justine Mrosak[1], MD; Swaminathan Kandaswamy[2], PhD; Claire Stokes[3,4], MD, MPH; David Roth[5], MD, MSPH; Jenna Gorbatkin[2], MD; Ishaan Dave[2], MSPH; Scott Gillespie[2], MS, MSPH; Evan Orenstein[2,4], MD

[1]Hennepin Healthcare, Minneapolis, MN, United States

[2]Department of Pediatrics, Emory University School of Medicine, Atlanta, GA, United States

[3]Division of Pediatric Hematology/Oncology, Department of Pediatrics, Emory University School of Medicine, Atlanta, GA, United States

[4]Children's Healthcare of Atlanta, Atlanta, GA, United States

[5]Department of Pediatrics, University of Pittsburgh School of Medicine, Pittsburgh, PA, United States

Corresponding Author:
Justine Mrosak, MD
Hennepin Healthcare
701 Park Avenue
Minneapolis, MN, 55415
United States
Phone: 1 6128891839
Email: jmrosak@gmail.com

Abstract

Background: Clinical practice guidelines (CPGs) and associated order sets can help standardize patient care and lead to higher-value patient care. However, difficult access and poor usability of these order sets can result in lower use rates and reduce the CPGs' impact on clinical outcomes. At our institution, we identified multiple CPGs for general pediatrics admissions where the appropriate order set was used in <50% of eligible encounters, leading to decreased adoption of CPG recommendations.

Objective: We aimed to determine how integrating disease-specific order groups into a common general admission order set influences adoption of CPG-specific order bundles for patients meeting CPG inclusion criteria admitted to the general pediatrics service.

Methods: We integrated order bundles for asthma, heavy menstrual bleeding, musculoskeletal infection, migraine, and pneumonia into a common general pediatrics order set. We compared pre- and postimplementation order bundle use rates for eligible encounters at both an intervention and nonintervention site for integrated CPGs. We also assessed order bundle adoption for nonintegrated CPGs, including bronchiolitis, acute gastroenteritis, and croup. In a post hoc analysis of encounters without order bundle use, we compared the pre- and postintervention frequency of diagnostic uncertainty at the time of admission.

Results: CPG order bundle use rates for incorporated CPGs increased by +9.8% (from 629/856, 73.5% to 405/486, 83.3%) at the intervention site and by +5.1% (896/1351, 66.3% to 509/713, 71.4%) at the nonintervention site. Order bundle adoption for nonintegrated CPGs decreased from 84% (536/638) to 68.5% (148/216), driven primarily by decreases in bronchiolitis order bundle adoption in the setting of the COVID-19 pandemic. Diagnostic uncertainty was more common in admissions without CPG order bundle use after implementation (28/227, 12.3% vs 19/81, 23.4%).

Conclusions: The integration of CPG-specific order bundles into a general admission order set improved overall CPG adoption. However, integrating only some CPGs may reduce adoption of order bundles for excluded CPGs. Diagnostic uncertainty at the time of admission is likely an underrecognized barrier to guideline adherence that is not addressed by an integrated admission order set.

(JMIR Med Inform 2023;11:e42736) doi:10.2196/42736

KEYWORDS

clinical practice guideline; user-centered design; clinical decision support; diagnostic uncertainty; diagnostic; decision support; CPG; clinical guideline; order bundle

Introduction

Problem Description

Clinical practice guidelines (CPGs) are designed to help standardize and disseminate evidence-based practices for various disease processes. Implementation of CPGs has been shown to decrease variation in care delivery, reduce costs, and improve patient outcomes [1-3]. However, clinician adherence to CPGs in many contexts remains suboptimal, impeding the delivery of high-value patient care [4,5].

Clinical decision support (CDS) systems integrated into the electronic health record can address some of these barriers and improve CPG adoption [6]. For example, order bundles serve as the building blocks for comprehensive order sets, which allow physicians to place multiple evidence-based orders for a single diagnosis with a few keystrokes without having to search individually for each order. CPG-associated order sets aggregate CPG-recommended therapies into a single order set, reducing the cognitive and physical work burden on clinicians to follow guidelines [6-9]. The use of CPG-associated order sets has improved outcomes in sepsis, pneumonia, and many other diseases [6,7,10,11].

Despite this evidence, we found that many CPG-associated order sets were used in <50% of eligible encounters in our own health system. The lack of order set adoption leads to less penetration of CPGs into clinical practice and reduces CPG impact on outcomes.

Available Knowledge

Research into guideline nonadherence thus far has demonstrated that barriers to guideline adoption are often context-specific and difficult to generalize to different settings [5]. In a systematic review of clinician surveys investigating potential barriers to guideline adherence, commonly identified barriers included lack of awareness or familiarity, lack of agreement or outcome expectancy, inertia from previous practice, and existing external barriers [5].

Order bundles embedded into larger order sets can be a powerful tool in improving CPG adoption [6,12]. For example, Munasinghe et al [10] incorporated multiple CPG order bundles into admission order sets and demonstrated improved adoption. However, unintended consequences, including increased physical and cognitive workload, can also result when these order sets demonstrate poor usability or are implemented at the wrong time in the clinician workflow [4,8]. Insufficient customization of order set content, mismatches between technology and human practices, and inadequate maintenance and modification of order sets have all been shown to contribute to order set nonuse and subsequent exposure to potential medical errors [4,8,9].

Furthermore, diagnostic uncertainty and increasing patient complexity also contribute to guideline nonadherence, which may be appropriate in certain contexts [13-15]. Single-diagnosis guidelines may be too simplistic to apply to the majority of patients [16]. Diagnosis codes have proven to be a poor marker of these barriers [14,15,17], and a lack of clear definitions continues to make them difficult to measure [14].

It remains unknown what CDS designs best address these barriers and most improve CPG adherence.

Rationale/Specific Aims

Our purpose for this project was to provide higher-value care through improved adherence to evidence-based CPGs at our institution. In preliminary data described elsewhere [18], we found that common reasons among local frontline providers for not adopting CPG order sets in eligible populations included lack of awareness (32%) and forgetting to use the stand-alone CPG-specific order set (20%). We therefore implemented a new CDS system in our admission process in the form of embedded CPG order bundles integrated into the general pediatrics admission order set that were identical to CPG-specific order bundles in the existing stand-alone CPG-associated order sets. Our primary aim was to increase CPG-specific order bundle use for eligible patients admitted on the general pediatrics service by 20% from July 2019 to May 2021. The primary aim was determined by the project stakeholder team to likely be an achievable improvement based on initial data that many CPGs demonstrated an <50% adherence rate, as well as an improvement that would justify the anticipated effort to complete the project. Secondary aims included determining if there were differences in order bundle use between specific CPGs at the intervention site and comparing CPG order bundle adherence between the intervention site and another hospital within the same health system where the intervention was not implemented.

Methods

Context/Setting

This study was performed on the general pediatrics service in an academic urban children's hospital within a 3-hospital, 638-bed pediatric health system serving the greater Atlanta, Georgia, area. Over 90 pediatric and family medicine residents rotate through the general pediatrics service each year and are overseen by 16 pediatric hospital medicine faculty members at the intervention hospital. Our institution uses an official electronic health record supplied by Epic Systems. Currently, there are over 20 general pediatric-specific CPGs customized to local workflows that are available for use through the institutional intranet. Prior to the intervention, 15 of the CPGs had their own stand-alone order set to facilitate adherence.

The intervention was implemented at 1 of 3 freestanding children's hospitals; this was an academic tertiary care center staffed primarily by resident teams with pediatric hospital medicine attendings. The nonintervention site was a hybrid community-academic hospital primarily staffed directly by pediatric hospital medicine attendings within the same health system.

Intervention

Planning the Intervention

Stakeholders included representatives from the Pediatric Hospital Medicine service, the Department of Clinical Effectiveness, and the Department of Quality and Safety, as well as a clinical informaticist, a human factors engineer, a

medical student, and a quality improvement methodology expert. This team met formally multiple times in the planning stages of the project and while formal problem analysis was underway before the intervention.

Preintervention problem analysis has been previously described [18]. Briefly, we identified patients eligible for a CPG order set for whom it was not ordered; we contacted the admitting provider within 2 weeks, inquired about reasons for CPG nonuse from a predefined list (we also added categories as needed), and asked for narrative comments. Based on these results, we created a Pareto chart that identified the most common barriers to CPG order set use: (1) lack of awareness or forgetting to use the CPG (2), eligibility for multiple CPG order sets at the time of admission, and (3) use of a similarly named order set that was not the intended CPG order set.

The Intervention

CPG-specific order bundles were integrated into the general pediatrics admission order set for better visibility and more efficient usability. Order bundles for 6 CPGs were chosen for the intervention, including asthma, complicated pneumonia, heavy menstrual bleeding, migraine, musculoskeletal infection, and uncomplicated pneumonia. These incorporated CPGs were chosen because they either (1) demonstrate low guideline order set use, (2) are very common, or (3) represent important improvement areas in antimicrobial stewardship. Order bundles were added to a section titled "Common Guidelines and Pathways—General Pediatrics" (Figure 1).

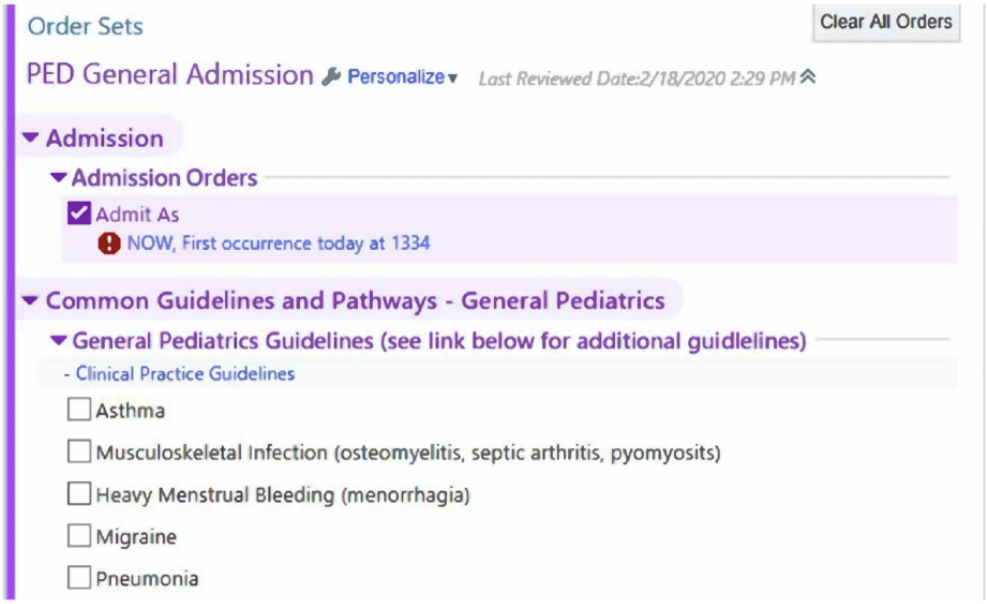
Orders in each CPG order bundle were identical to the existing stand-alone CPG-associated order sets. Within each order bundle, embedded hyperlinks referenced the published CPG and relevant literature from which recommendations were made and referenced common target disease pathogens for bundles that recommended empiric first-line antibiotics. For patients that qualified for multiple CPG order sets, the integrated order set also allowed for the selection of multiple relevant order bundles within the order set.

Prior to implementation of the integrated order set into a live production environment, formative usability testing and summative usability testing were both completed. Usability testing aimed to test the effectiveness and iteratively improve the intervention in a simulated environment. Results are reported elsewhere [18].

To address the barrier of similarly named but non-CPG order sets being used accidentally, we identified order set "mimics" by searching the system with common clinical synonyms for each CPG. Similarly named order sets were retired from the production environment after obtaining approval from both the order set owners and the corresponding CPG owners. In total, 9 mimics were identified, and all of these were subsequently retired after owner approval. Additionally, all relevant CPG-associated order sets were reviewed and updated, if necessary, for both naming consistency and related search terms.

After the integrated order set was implemented into a live production environment, an update outlining the new CDS tool and its capabilities and expected use was emailed to all current and incoming residents to review and presented at the weekly resident educational conference.

Figure 1. Integrated general pediatrics admission order set with clinical practice guideline order bundles.



Study of the Intervention and Measures

All patients aged 0 to 21 years who were admitted to the general pediatrics service and met the eligibility criteria for any one of the incorporated CPG order sets based on preexisting computable population definitions were included in this study. This study used a quasi-experimental design, analyzing pre- and postintervention CPG order set adherence at both the intervention site and a nonintervention site. Our primary exposure was intervention period, with the preintervention period defined as July 1, 2019, to June 3, 2020, and the postintervention period as June 4, 2020, to May 28, 2021, as the integrated order set went live on June 4, 2020, at the intervention site.

Our primary outcome was the proportion of appropriate CPG order set use for eligible patients at the time of admission. To evaluate the impact of our intervention, we adopted existing automated queries to assess whether the clinician used the appropriate CPG order set, a wrong but similarly named order set, or the available general pediatrics admission order set. We also reviewed whether the "CPG guideline initiation order" was signed, which is a prechecked order in all our included CPG order sets. Through the query, demonstrated use of the CPG-associated order bundle and the presence of the guideline initiation order were assumed to represent appropriate guideline order set use. Encounters where the clinician appeared to use the appropriate guideline order set or bundle but the guideline initiation order was absent were manually chart reviewed to confirm appropriate order set use. All encounters that appeared eligible but where the clinician did not appear to demonstrate appropriate guideline order set use were manually chart reviewed to ensure CPG eligibility at admission throughout the study period. Eligibility was based on defined eligibility criteria in each published CPG. All manual chart review was completed by a pediatric hospital medicine fellow using both the Epic electronic health record and the Phrase Health system.

The pre- and postintervention proportion of CPG-eligible admissions for which the CPG order set or bundle were used was compared at the intervention hospital, where the integrated order set was implemented into the production environment. Data in this context were considered our "intervention cohort." The proportion of appropriate CPG order set use for eligible patients was also compared in the same study period at the nonintervention hospital within the same health system, where the integrated order set was not implemented. This hospital uses the same CPGs and associated order sets and serves a similar patient population in the greater Atlanta area; it was thus considered our "nonintervention cohort." The purpose of having both intervention and nonintervention cohorts in this study was to better assess whether the observed outcomes were directly related to our intervention rather than secular trends.

In this intervention, there was concern that surfacing some guidelines in the integrated order set but not others could lead to the unintended consequence of reducing order set use of CPGs that were not included in the intervention. Therefore, pre- and postintervention use of CPG order sets that were not initially included in the integrated admission order set (acute gastroenteritis, croup, bronchiolitis) was assessed as a "balancing cohort."

Evaluation of Diagnostic Uncertainty

As our intervention was created to address lack of knowledge and awareness of guidelines, we hypothesized that it would not address diagnostic uncertainty, an underrecognized barrier that may influence order set adoption. In a post hoc analysis, we therefore aimed to evaluate the presence of diagnostic uncertainty at the time of admission to determine if this barrier accounted for a larger proportion of CPG order bundle nonuse after the intervention.

All eligible encounters where the associated CPG order set was not used were manually chart reviewed to assess the presence of diagnostic uncertainty throughout the entire study period. Diagnostic uncertainty was defined based on an algorithm (Multimedia Appendix 1) adapted from the approach of Bhise et al [15] to measuring diagnostic uncertainty in primary care. In the algorithm, encounters needed to include direct or indirect markers of uncertainty in documentation, initial definitive treatment had to have been withheld while awaiting further diagnostic workup or observation, and an operational definition of diagnostic uncertainty had to be met. Two members of the research team, a pediatric hospital medicine fellow and a pediatric resident, completed the manual chart review based on information available in the initial history and physical documentation and reported the presence or absence of diagnostic uncertainty. Interrater reliability was assessed to confirm reliability between the 2 researchers' assessments. After chart review, the number of eligible encounters where the CPG-associated order set was not used that demonstrated the presence of diagnostic uncertainty was compared before and after implementation to determine the change in proportion after the intervention.

Analysis

Data were summarized using counts and percentages by site (intervention and nonintervention), period (pre- and postintervention), and guideline (eg, asthma or heavy menstrual bleeding). Binary logistic regression was used to analyze overall and by-guideline associations between use of appropriate CPG-specific order bundle (yes vs no) and period (pre- and postintervention) across sites via statistical interactions. We further ran binary logistic regression models evaluating the association between use of appropriate CPG-specific order bundle and period in the balancing cohort and relevant guidelines. Results are presented as contingency tables with odds ratios (ORs), 95% CIs, and corresponding *P* values. All analyses were conducted using SAS (version 9.4; SAS Institute), and significance was assessed at the .05 level. Percent adherence for eligible encounters by month for the intervention cohort was tracked and plotted on a statistical process *P* chart with annotations for the order set clean up and integrated order set go-live interventions.

Ethical Considerations

This study was deemed by the Children's Healthcare of Atlanta Institutional Review Board to be nonhuman-subjects research as a quality improvement study (STUDY00000367).

Results

The integrated order set went live on June 4, 2020. From January 1, 2019, to May 28, 2021, a total of 1664 encounters were identified as eligible for a CPG order set based on preexisting computable population definitions. Of these encounters, 1052 were preimplementation (Figure 2) and 612 were postimplementation (Figure 3).

Figure 2. Preintervention admission encounters for the intervention cohort. CPG: clinical practice guideline.

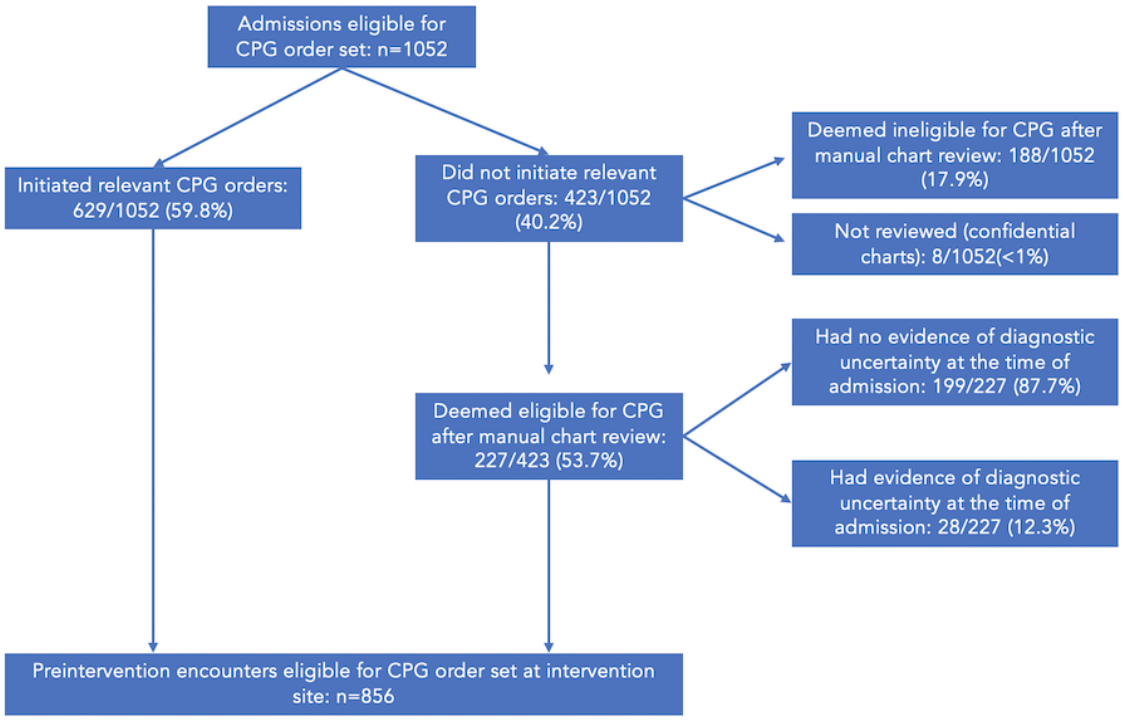
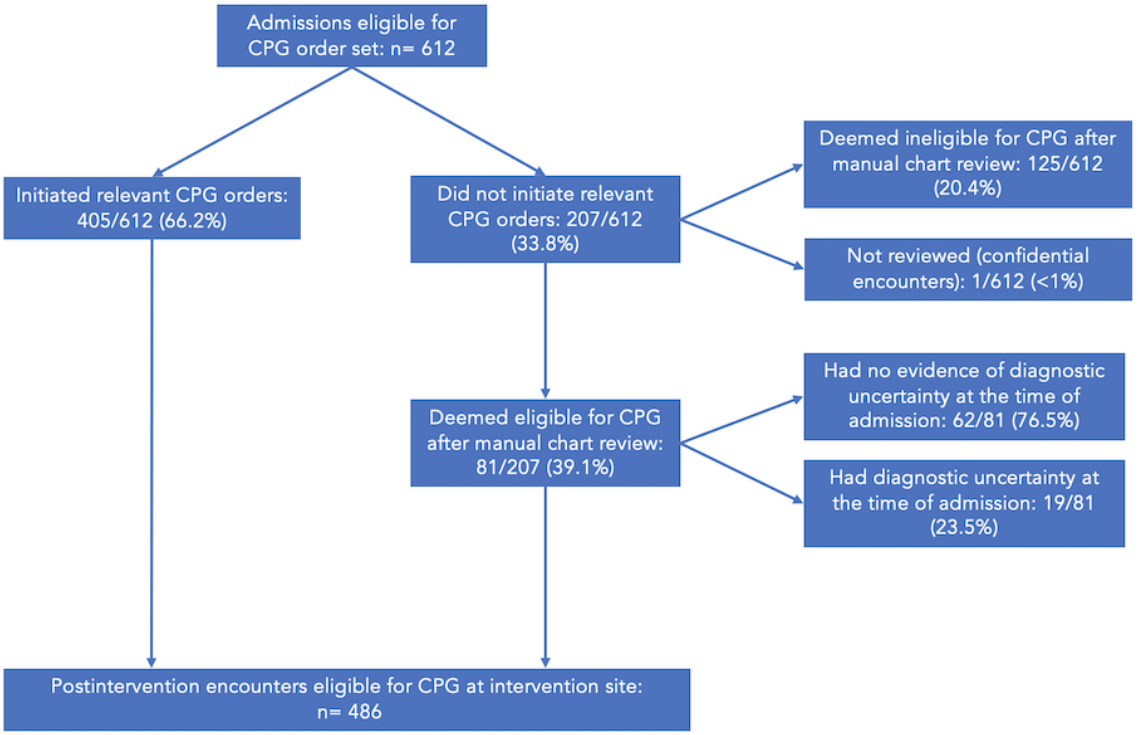


Figure 3. Postintervention admission encounters for the intervention cohort. CPG: clinical practice guideline.



The number of encounters was unbalanced, partially due to lower-than-average admission volumes as a result of the COVID-19 pandemic. We manually reviewed all encounters that appeared eligible by computable definitions for a CPG order bundle where the CPG order bundle was not used: 423/1052 (40.2%) encounters before the intervention and 207/612 (33.8%) after the intervention. Of the encounters that were reviewed, 188/1052 (17.9%) before the intervention and 125/612 (20.4%) after the intervention were excluded by manual review for not meeting eligibility criteria. Overall rates of exclusion were similar when comparing the difference before and after the intervention (17.9% before and 20.4% after the intervention; 95% CI –6.62% to 1.51%, *P*=.22).

CPG order set use rates for included CPGs were tracked over time (Figure 4).

The trend in monthly adherence was positive following implementation and demonstrated special cause variation

beginning in August 2020, 8 weeks after the integrated order set went live. The rate of order set use at the intervention site for integrated CPGs increased from 73.5% before the intervention to 83.3% after the intervention (OR 1.80, 95% CI 1.36-2.39). Order set use rate at the nonintervention site, where the integrated order set was not implemented but mimics were also deleted, increased from 66.3% to 71.4% during the same study period (OR 1.27, 95% CI 1.04-1.54). Of note, this increase in the nonintervention cohort appeared driven by musculoskeletal infection (OR 2.84, 95% CI 1.49-5.40) and asthma (OR 2.15, 95% CI 1.22-3.79), as seen in Table 1. When comparing ORs between the intervention and nonintervention cohorts, the intervention cohort had significantly improved order set use from before to after the intervention relative to the nonintervention cohort (intervention OR 1.80 (95% CI 1.36-2.39) vs nonintervention OR 1.27 (95% CI 1.04-1.54; *P*=.045).

When broken down by disease-specific CPGs, all integrated CPGs showed positive adherence trends after implementation in the intervention cohort but with different effect sizes. Heavy menstrual bleeding and pneumonia had more improvement than musculoskeletal infection or migraine (Table 1). Adherence in asthma, for which the CPG order set has historically high use rates, remained excellent after the intervention (92.1%-95.5%; OR 1.81, 95% CI 0.95-3.45). Adoption of CPG order bundles that were not included in the integrated admission order set (including bronchiolitis, acute gastroenteritis, and croup) decreased from 84% to 68.5% following the intervention (OR 0.41, 95% CI 0.29-0.59). Of note, this was largely driven by bronchiolitis, where adoption changed from 86.9% to 75.7% after the intervention (OR 0.47, 95% CI 0.28-0.78), as seen in Table 2.

In a post hoc analysis, based on the observation that improvements were lower for musculoskeletal infection and migraine, we reviewed 308 eligible encounters where a CPG order bundle was not used to evaluate the presence of diagnostic uncertainty at admission. One reviewer (a pediatric hospital medicine fellow) completed manual chart review on all charts and a second (a pediatric resident) reviewed a random subsample of 50 encounters (16%), with interrater reliability measured by the Cohen κ (κ=0.73, *P*<.001). The proportion of eligible encounters where the CPG order set was not used that demonstrated diagnostic uncertainty increased from 12.3% (28/227) before implementation to 23.4% (19/81) after implementation (OR 2.18, 95% CI 1.12-4.16).

Figure 4. Statistical process control chart of percentage guideline order set adherence for eligible encounters at the intervention site between July 2019 to December 2020. OS: order set.

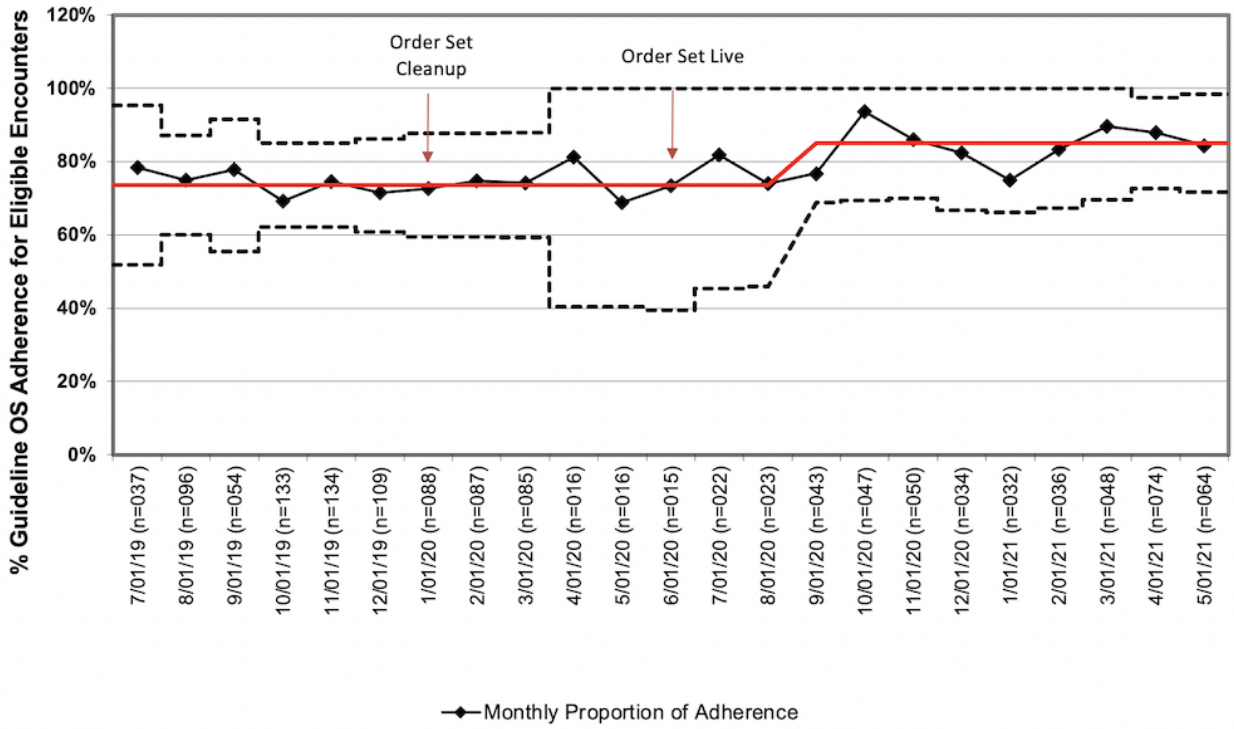


Table 1. Order set bundle use before and after implementation on June 4, 2020, in the intervention and nonintervention cohorts.

	Intervention cohort				Nonintervention cohort				Interaction *P* value
	No bundle, n (%)	Bundle, n (%)	OR[a] (95% CI)	*P* value	No bundle, n (%)	Bundle, n (%)	OR (95% CI)	*P* value	
Overall[b]				<.001				.02	.04
Before	227 (26.5)	629 (73.5)	Reference		455 (33.7)	896 (66.3)	Reference		
After	81 (16.7)	405 (83.3)	1.80 (1.36-2.39)		204 (28.6)	509 (71.4)	1.27 (1.04-1.54)		
Asthma[c]				.07				.01	.69
Before	39 (7.9)	454 (92.1)	Reference		63 (10.6)	529 (89.4)	Reference		
After	13 (4.5)	274 (95.5)	1.81 (0.95-3.45)		16 (5.2)	289 (94.8)	2.15 (1.22-3.79)		
Heavy menstrual bleeding[d]				.01				.24	.01
Before	8 (42.1)	11 (57.9)	Reference		2 (6.7)	28 (93.3)	Reference		
After	5 (11.6)	38 (88.4)	5.53 (1.50-20.35)		8 (16.0)	42 (84.0)	0.38 (0.07-1.90)		
Musculoskeletal infection[e]				.30				.001	.48
Before	22 (75.9)	7 (24.1)	Reference		55 (67.1)	27 (32.9)	Reference		
After	21 (63.6)	12 (36.4)	1.80 (0.59-5.44)		33 (41.8)	46 (58.2)	2.84 (1.49-5.40)		
Migraine[f]				.77				.08	.26
Before	25 (30.1)	58 (69.9)	Reference		75 (37.1)	127 (62.9)	Reference		
After	19 (27.9)	49 (72.1)	1.11 (0.55-2.26)		81 (46.0)	95 (54.0)	0.69 (0.46-1.05)		
Complicated pneumonia[g]				.27				.19	.09
Before	7 (87.5)	1 (12.5)	Reference		28 (65.1)	15 (34.9)	Reference		
After	5 (62.5)	3 (37.5)	4.20 (0.33-53.11)		8 (88.9)	1 (11.1)	0.23 (0.03-2.05)		
Uncomplicated pneumonia[h]				.03				.48	.03
Before	126 (56.3)	98 (43.7)	Reference		232 (57.7)	170 (42.3)	Reference		
After	18 (38.3)	29 (61.7)	2.07 (1.09-3.95)		58 (61.7)	36 (38.3)	0.85 (0.53-1.34)		

[a]OR: odds ratio.

[b]Intervention cohort: n=1342; nonintervention cohort: n=2064.

[c]Intervention cohort: n=780; nonintervention cohort: n=897.

[d]Intervention cohort: n=62; nonintervention cohort: n=80.

[e]Intervention cohort: n=62; nonintervention cohort: n=161.

[f]Intervention cohort: n=151; nonintervention cohort: n=378.

[g]Intervention cohort: n=16; nonintervention cohort: n=52.

[h]Intervention cohort: n=271; nonintervention cohort: n=496.

Table 2. Pre- and postimplementation (before and on or after June 4, 2020, respectively) order set bundle use in the balancing cohort.

	No bundle, n (%)	Bundle, n (%)	Odds ratio (95% CI)	*P* value
Overall (n=854)				
Preimplementation	102 (16)	536 (84)	Reference	
Postimplementation	68 (31.5)	148 (68.5)	0.41 (0.29-0.59)	<.001
Bronchiolitis (n=563)				
Preimplementation	59 (13.1)	393 (86.9)	Reference	
Postimplementation	27 (24.3)	84 (75.7)	0.47 (0.28-0.78)	.004
Acute gastroenteritis (n=109)				
Preimplementation	33 (63.5)	19 (36.5)	Reference	
Postimplementation	35 (61.4)	22 (38.6)	1.09 (0.50-2.37)	.83
Croup (n=184)				
Preimplementation	10 (7.5)	124 (92.5)	Reference	
Postimplementation	6 (12.5)	42 (87.5)	0.57 (0.19-1.65)	.29

Discussion

Summary

The integration of CPG order bundles into a general pediatric admission order set improved CPG adoption in a stand-alone academic pediatric hospital compared to a control hospital within the same health system. In a post hoc analysis, the disease processes with lower diagnostic uncertainty at the time of admission saw the greatest improvement from this intervention.

Interpretation

CPG adoption improved both in relation to preintervention encounters at the same hospital and in relation to encounters at a similar hospital within the same institution where the integrated order set was not released. This suggests that the increase in CPG adherence was directly related to the implementation of the integrated order set at the study site. While CPG adherence also significantly improved at the nonintervention hospital, the improvement seen at the intervention site was significantly more than that at the nonintervention site. Additionally, improvement was only seen for 2 of the 6 guidelines (asthma and musculoskeletal infection) at the nonintervention site, compared to all 6 guidelines showing a trend toward improvement at the intervention site. This change may reflect the removal of known CPG order set “mimics” at both locations prior to the integrated order set implementation, as this was identified as a barrier to CPG adherence in prior work. While some CPGs demonstrated improvements of close to 20%, overall improvement did not meet our initial primary aim of 20% increased adherence after the intervention. This is likely due to finding a higher than anticipated preintervention overall adherence rate, largely driven by the CPG for asthma, which has historically high adherence rates.

Nonincorporated CPGs demonstrated a reduction in order set use following implementation of the integrated order set. This finding was largely attributable to a decrease in bronchiolitis guideline adherence. The timing of this intervention, in June 2020, correlated to multiple surges of SARS-CoV-2 infections. We are unable to distinguish whether the reduction in adherence was due to our intervention or the change in the management of respiratory infections during this time. Future studies that incorporate a more comprehensive list of CPGs may elucidate how this decision support design affects nonincorporated CPGs.

The presence of diagnostic uncertainty was not initially identified as a primary barrier to guideline adherence based on frontline clinician queries during this study [18]. In our analysis, the proportion of eligible encounters without CPG adherence that demonstrated diagnostic uncertainty increased following the implementation of the integrated order set. Our intervention addressed other drivers, but not diagnostic uncertainty, which may explain a higher fraction of diagnostic uncertainty in encounters without a CPG order bundle after implementation. Alternative designs that account for the change in diagnostic certainty across a hospitalization may demonstrate a better job of improving CPG adherence.

While previous literature has shown that CPGs and associated order sets can successfully decrease variation in care delivery and improve patient outcomes [19-22], the context in which decision support is aligned into the workflow remains of utmost importance for the success of these interventions [4]. Combining alerts with order sets has been shown to have success in specific contexts [23,24] but risks generating alert fatigue and requires considerable disease-specific logic behind the alert. Alternatively, automating order suggestions through machine learned patterns of order use was shown to influence ordering behavior [4,25] but may not reflect evidence-based recommendations and can be resource intensive.

Furthermore, increasing patient medical complexity and diagnostic uncertainty leads to a workflow mismatch when coupled with single diagnoses or simplified guidelines. This gap was largely underrecognized by clinicians when self-reporting barriers to guideline adherence [18] and likely requires decision support in a different context or format to overcome than admission order sets. Mehta et al [16] sought to integrate CDS into documentation workflows through problem-oriented templates aimed at improving documentation

for patients with multiple problems and to provide greater evidence-based prompts and organization. While demonstrating potential to provide recommendations during the documentation process, potentially a better context to address diagnostic uncertainty, clinicians are still called upon to label diagnoses for their patients at a point in time when this may still be unclear. Further research into the most effective format and context for CDS to address diagnostic uncertainty is needed.

Limitations

This study has several limitations. First, results may not be generalizable, as this was a multisite, single-system study focused on a single service. Different contexts, organizational cultures, and electronic health record vendors could affect the feasibility and impact of this intervention. Second, due to the COVID-19 pandemic, hospital admission volumes were significantly lower, and admissions consisted of fewer respiratory illnesses in the postimplementation period, potentially creating pre- and postintervention cohorts that were less similar. Additionally, due to time and resource constraints, a single chart reviewer was used for charts flagged as nonadherent to confirm eligibility. Despite this limitation, pre- and postimplementation rates of exclusion were similar, suggesting this did not have a significant influence on results. Lastly, in a post hoc analysis of diagnostic uncertainty, only one reviewer reviewed all nonadherent charts to determine the presence of uncertainty. The development of an algorithm for uncertainty and an assessment of interrater reliability for a subset of charts attempted to address this limitation and minimize subjectivity.

Conclusion

The integration of CPG-specific order bundles into a general pediatrics admission order set improved overall CPG adoption by addressing the most commonly reported barriers to CPG adherence by clinicians. Further improvement in guideline adherence could be seen with integration of a more comprehensive list of available guidelines for a particular service. Diagnostic uncertainty at the time of admission is likely an underrecognized barrier to guideline adherence that is not fully addressed with an integrated admission order set. Further work is needed to determine the impact of an integrated admission order set on clinical outcomes and what types of clinical decision support could better address the presence of diagnostic uncertainty.

Acknowledgments

We would like to acknowledge Christy Bryant for her assistance in building the redesigned order set as well as the pediatric residents and attendings who participated in all stages of this study. JM, SK, SG, and ID's participation was supported by the Emory Department of Pediatrics and Children's Healthcare of Atlanta through the Warshaw Fellow Research Award.

Authors' Contributions

JM, SK, CS, DR, and EO conceptualized and designed the project. JM performed manual chart review for nonadherence. JM and JG created the algorithm for diagnostic uncertainty and performed chart review for diagnostic uncertainty. JG structured and wrote Multimedia Appendix 1 with input from JM. ID and SG performed statistical analysis and contributed to interpretation of results. All authors discussed the results and conclusions of the study. JM wrote the manuscript with contribution and input from all authors.

Conflicts of Interest

EO is a cofounder of and has equity in Phrase Health, a clinical decision support analytics company. He receives no direct revenue from this relationship. No other authors declare any conflicts.

Multimedia Appendix 1
Overview of algorithm to evaluate for diagnostic uncertainty with case examples.
[DOCX File , 76 KB - medinform_v11i1e42736_app1.docx]

References

1. Lion KC, Wright DR, Spencer S, Zhou C, Del Beccaro M, Mangione-Smith R. Standardized clinical pathways for hospitalized children and outcomes. Pediatrics 2016 Apr;137(4):e20151202 [FREE Full text] [doi: 10.1542/peds.2015-1202] [Medline: 27002007]
2. Kasmire KE, Hoppa EC, Patel PP, Boch KN, Sacco T, Waynik IY. Reducing invasive care for low-risk febrile infants through implementation of a clinical pathway. Pediatrics 2019 Mar;143(3):e20181610. [doi: 10.1542/peds.2018-1610] [Medline: 30728272]
3. Nkoy F, Fassl B, Stone B, Uchida DA, Johnson J, Reynolds C, et al. Improving pediatric asthma care and outcomes across multiple hospitals. Pediatrics 2015 Dec;136(6):e1602-e1610. [doi: 10.1542/peds.2015-0285] [Medline: 26527553]
4. Li RC, Wang JK, Sharp C, Chen JH. When order sets do not align with clinician workflow: assessing practice patterns in the electronic health record. BMJ Qual Saf 2019 Dec;28(12):987-996 [FREE Full text] [doi: 10.1136/bmjqs-2018-008968] [Medline: 31164486]

5. Cabana MD, Rand CS, Powe NR, Wu AW, Wilson MH, Abboud PA, et al. Why don't physicians follow clinical practice guidelines? A framework for improvement. JAMA 1999 Oct 20;282(15):1458-1465. [doi: 10.1001/jama.282.15.1458] [Medline: 10535437]
6. Kaiser SV, Lam R, Cabana MD, Bekmezian A, Bardach NS, Auerbach A, Pediatric Research in Inpatient Settings (PRIS) Network. Best practices in implementing inpatient pediatric asthma pathways: a qualitative study. J Asthma 2020 Jul;57(7):744-754. [doi: 10.1080/02770903.2019.1606237] [Medline: 31020879]
7. Bright TJ, Wong A, Dhurjati R, Bristow E, Bastian L, Coeytaux RR, et al. Effect of clinical decision-support systems: a systematic review. Ann Intern Med 2012 Jul 03;157(1):29-43 [FREE Full text] [doi: 10.7326/0003-4819-157-1-201207030-00450] [Medline: 22751758]
8. Zhang Y, Padman R, Levin JE. Paving the COWpath: data-driven design of pediatric order sets. J Am Med Inform Assoc 2014 Oct;21(e2):e304-e311 [FREE Full text] [doi: 10.1136/amiajnl-2013-002316] [Medline: 24674844]
9. Gartner D, Zhang Y, Padman R. Cognitive workload reduction in hospital information systems : Decision support for order set optimization. Health Care Manag Sci 2018 Jun;21(2):224-243. [doi: 10.1007/s10729-017-9406-6] [Medline: 28551859]
10. Munasinghe RL, Arsene C, Abraham TK, Zidan M, Siddique M. Improving the utilization of admission order sets in a computerized physician order entry system by integrating modular disease specific order subsets into a general medicine admission order set. J Am Med Inform Assoc 2011 May 01;18(3):322-326 [FREE Full text] [doi: 10.1136/amiajnl-2010-000066] [Medline: 21422099]
11. Goldszer RC, Ratzan K, Csete M, Nanes N, Love C, Cubeddu LX, et al. Impact of order set use on outcome of patients with sepsis. Appl Inform 2017 Jan 5;4(1):1-6 [FREE Full text] [doi: 10.1186/s40535-016-0033-y]
12. Fishbane S, Niederman MS, Daly C, Magin A, Kawabata M, de Corla-Souza A, et al. The impact of standardized order sets and intensive clinical case management on outcomes in community-acquired pneumonia. Arch Intern Med 2007;167(15):1664-1669. [doi: 10.1001/archinte.167.15.1664] [Medline: 17698690]
13. Whaley LE, Businger AC, Dempsey PP, Linder JA. Visit complexity, diagnostic uncertainty, and antibiotic prescribing for acute cough in primary care: a retrospective study. BMC Fam Pract 2013 Aug 19;14:120 [FREE Full text] [doi: 10.1186/1471-2296-14-120] [Medline: 23957228]
14. Bhise V, Rajan SS, Sittig DF, Morgan RO, Chaudhary P, Singh H. Defining and measuring diagnostic uncertainty in medicine: a systematic review. J Gen Intern Med 2018 Jan;33(1):103-115 [FREE Full text] [doi: 10.1007/s11606-017-4164-1] [Medline: 28936618]
15. Bhise V, Rajan SS, Sittig DF, Vaghani V, Morgan RO, Khanna A, et al. Electronic health record reviews to measure diagnostic uncertainty in primary care. J Eval Clin Pract 2018 Jun;24(3):545-551. [doi: 10.1111/jep.12912] [Medline: 29675888]
16. Mehta R, Radhakrishnan NS, Warring CD, Jain A, Fuentes J, Dolganiuc A, et al. The use of evidence-based, problem-oriented templates as a clinical decision support in an inpatient electronic health record system. Appl Clin Inform 2016 Aug 17;7(3):790-802 [FREE Full text] [doi: 10.4338/ACI-2015-11-RA-0164] [Medline: 27530268]
17. Marshall TL, Ipsaro AJ, Le M, Sump C, Darrell H, Mapes KG, et al. Increasing physician reporting of diagnostic learning opportunities. Pediatrics 2021 Jan;147(1):e20192400 [FREE Full text] [doi: 10.1542/peds.2019-2400] [Medline: 33268395]
18. Mrosak J, Kandaswamy S, Stokes C, Roth D, Dave I, Gillespie S, et al. The influence of integrating clinical practice guideline order bundles into a general admission order set on guideline adoption. JAMIA Open 2021 Oct;4(4):ooab087 [FREE Full text] [doi: 10.1093/jamiaopen/ooab087] [Medline: 34632324]
19. Leighton H, Kianfar H, Serynek S, Kerwin T. Effect of an electronic ordering system on adherence to the American College of Cardiology/American Heart Association guidelines for cardiac monitoring. Crit Pathw Cardiol 2013 Mar;12(1):6-8. [doi: 10.1097/HPC.0b013e318270787c] [Medline: 23411601]
20. Kitchlu A, Abdelshaheed T, Tullis E, Gupta S. Gaps in the inpatient management of chronic obstructive pulmonary disease exacerbation and impact of an evidence-based order set. Can Respir J 2015;22(3):157-162 [FREE Full text] [doi: 10.1155/2015/587026] [Medline: 25886627]
21. Bartlett KW, Parente VM, Morales V, Hauser J, McLean HS. Improving the efficiency of care for pediatric patients hospitalized with asthma. Hosp Pediatr 2017 Jan;7(1):31-38. [doi: 10.1542/hpeds.2016-0108] [Medline: 27932381]
22. Joyner Blair AM, Hamilton BK, Spurlock A. Evaluating an order set for improvement of quality outcomes in diabetic ketoacidosis. Adv Emerg Nurs J 2018;40(1):59-72. [doi: 10.1097/TME.0000000000000178] [Medline: 29384776]
23. Berger RP, Saladino RA, Fromkin J, Heineman E, Suresh S, McGinn T. Development of an electronic medical record-based child physical abuse alert system. J Am Med Inform Assoc 2018 Feb 01;25(2):142-149 [FREE Full text] [doi: 10.1093/jamia/ocx063] [Medline: 28641385]
24. Suresh S, Saladino RA, Fromkin J, Heineman E, McGinn T, Richichi R, et al. Integration of physical abuse clinical decision support into the electronic health record at a Tertiary Care Children's Hospital. J Am Med Inform Assoc 2018 Jul 01;25(7):833-840 [FREE Full text] [doi: 10.1093/jamia/ocy025] [Medline: 29659856]
25. Wang JK, Hom J, Balasubramanian S, Schuler A, Shah NH, Goldstein MK, et al. An evaluation of clinical order patterns machine-learned from clinician cohorts stratified by patient mortality outcomes. J Biomed Inform 2018 Oct;86:109-119 [FREE Full text] [doi: 10.1016/j.jbi.2018.09.005] [Medline: 30195660]

Abbreviations

CDS: clinical decision support
CPG: clinical practice guidelines
OR: odds ratio

Edited by C Lovis; submitted 15.09.22; peer-reviewed by J Ray; comments to author 10.11.22; revised version received 30.11.22; accepted 01.12.22; published 21.03.23.

Please cite as:
Mrosak J, Kandaswamy S, Stokes C, Roth D, Gorbatkin J, Dave I, Gillespie S, Orenstein E
The Effect of Implementation of Guideline Order Bundles Into a General Admission Order Set on Clinical Practice Guideline Adoption: Quasi-Experimental Study
JMIR Med Inform 2023;11:e42736
URL: https://medinform.jmir.org/2023/1/e42736
doi:10.2196/42736
PMID:36943348

An Extract-Transform-Load Process Design for the Incremental Loading of German Real-World Data Based on FHIR and OMOP CDM: Algorithm Development and Validation

Elisa Henke[*], MSc; Yuan Peng[*], MSc; Ines Reinecke, MSc; Michéle Zoch, Dipl Wi Inf; Martin Sedlmayr, Dr Rer Nat, Prof Dr; Franziska Bathelt, Dr Rer Nat

Institute for Medical Informatics and Biometry, Carl Gustav Carus Faculty of Medicine, Technische Universität Dresden, Dresden, Saxony, Germany
[*]these authors contributed equally

Corresponding Author:
Elisa Henke, MSc

Abstract

Background: In the Medical Informatics in Research and Care in University Medicine (MIRACUM) consortium, an IT-based clinical trial recruitment support system was developed based on the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM). Currently, OMOP CDM is populated with German Fast Healthcare Interoperability Resources (FHIR) using an Extract-Transform-Load (ETL) process, which was designed as a bulk load. However, the computational effort that comes with an everyday full load is not efficient for daily recruitment.

Objective: The aim of this study is to extend our existing ETL process with the option of incremental loading to efficiently support daily updated data.

Methods: Based on our existing bulk ETL process, we performed an analysis to determine the requirements of incremental loading. Furthermore, a literature review was conducted to identify adaptable approaches. Based on this, we implemented three methods to integrate incremental loading into our ETL process. Lastly, a test suite was defined to evaluate the incremental loading for data correctness and performance compared to bulk loading.

Results: The resulting ETL process supports bulk and incremental loading. Performance tests show that the incremental load took 87.5% less execution time than the bulk load (2.12 min compared to 17.07 min) related to changes of 1 day, while no data differences occurred in OMOP CDM.

Conclusions: Since incremental loading is more efficient than a daily bulk load and both loading options result in the same amount of data, we recommend using bulk load for an initial load and switching to incremental load for daily updates. The resulting incremental ETL logic can be applied internationally since it is not restricted to German FHIR profiles.

(JMIR Med Inform 2023;11:e47310) doi:10.2196/47310

KEYWORDS

ETL; incremental loading; OMOP CDM; FHIR; interoperability; Extract-Transform-Load; Observational Medical Outcomes Partnership Common Data Model; Fast Healthcare Interoperability Resources

Introduction

Background and Significance

Randomized controlled trials are the gold standard to “measure the effectiveness of a new intervention or treatment” [1]. However, randomized controlled trials are limited regarding the representative number of persons included and, therefore, restricted in their external generalizability. To gain more unbiased evidence, observational studies focus on real-world data from large heterogeneous populations.

To support observational research, we already provide a transferable Extract-Transform-Load (ETL) process [2] to transform German real-world data to the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) [3] provided by Observational Health Data Sciences and Informatics (OHDSI) [4], which supports the possibilities for multicentric and even international studies. Due to the heterogeneity of the structure and content of the data from the data integration centers within the Medical Informatics Initiative Germany (MI-I) [5], the Health Level 7 (HL7) [6] Fast Healthcare Interoperability Resources (FHIR) communication standard was specified among all German university hospitals. Consequently, we used FHIR as the source for our ETL process. The FHIR specification is given by the core data set of the MI-I [7]. FHIR resources can be read from an FHIR Gateway [8] (PostgreSQL database) or FHIR Server (eg, HAPI [9] or Blaze [10]). As the target of our ETL process, we used OMOP CDM v5.3.1 [11]. The implementation of the ETL process was done using the open source framework Java SpringBatch [12]. Our

ETL process has been implemented in accordance with the default assumption as described in the Book of OHDSI [13], where the OHDSI community defines the ETL process as a full load to transfer data from source to target systems. This approach is efficient for a dedicated study where data gets loaded once without any update afterward but inefficient if it comes to the need for updated data on a daily basis.

Latter is the case for the developments around the improvement and support of the recruitment process for clinical trials, which the Medical Informatics in Research and Care in University Medicine (MIRACUM) [14] consortium, as part of the MI-I funded by the German Federal Ministry of Education and Research, is working on. In this context, an IT-based clinical trial recruitment support system (CTRSS) based on OMOP CDM was implemented [15]. The CTRSS consists of a screening list for recruitment teams that provides potential candidates for clinical trials updated on a daily base. To enable the CTRSS to provide recruitment proposals, it is necessary to transform the data in FHIR format at each site from the 10 MIRACUM data integration centers into the standardized format of OMOP CDM. The procession of FHIR resources to OMOP CDM through our ETL process has already been successfully tested and integrated at all 10 German university hospitals of the MIRACUM consortium.

So far, our ETL process is restricted to a bulk load of FHIR resources to OMOP CDM. This implied that all FHIR resources are read from the source. To enable the CTRSS to provide daily recruitment proposals, our ETL process has to be executed every day as a full load. However, an everyday full load is not efficient because often only a small amount of source data has changed during loading periods, which results in unnecessary long execution times considering a full load for daily executions. Consequently, the computational effort that comes with the daily execution of the bulk load is not efficient in the context of the CTRSS.

Thus, a new approach is needed to only process FHIR resources that were created, updated, or deleted (CUD) since the last execution of the ETL process once an initial load has been executed. This loading option is known as incremental loading.

Objective

To keep the bulk load option for dedicated studies and still be performant toward daily changes in the source data, a combination of bulk load and incremental load is needed. To reduce the additional effort in implementing a second independent ETL process for incremental loading, it is our aim to extend our existing ETL process with the option of incremental loading. During our research, we focused on the following four research questions:

1. What requirements need to be considered when integrating incremental loading into our existing ETL process design?
2. What approaches already exist for incremental ETL processes?
3. How can the identified requirements from research question 1 be implemented in our existing ETL process design?
4. Does incremental loading provide an advantage over daily bulk loading?

Methods

Analysis of the Existing ETL FHIR-to-OMOP Process

To determine the requirements for integrating incremental loading into our existing ETL process design, we performed an impact analysis focusing on the whole ETL process as well as, in more detail, the three main components of it, namely, Reader, Processor, and Writer as presented by Peng et al [2]. Regarding the whole ETL process, the following 3 requirements were needed:

- Requirement A: It is necessary to provide the user with the ability to distinguish between bulk loading and incremental loading.
- Requirement B: For incremental loading, it is further essential that the Reader of the ETL process is able to detect changes in the source system and reads only CUD-FHIR resources on a daily basis.
- Requirement C: During the processing of updated and deleted FHIR resources, duplicates and obsolete data should be avoided in OMOP CDM to guarantee data correctness.

Considering the semantic mapping from FHIR MI-I Core Data Set (CDS) to OMOP CDM and the Writer of the ETL process as described by Peng et al [2], incremental loading has no impact on both. In summary, incremental loading requires an adjustment of the implementation of the Reader and Processor.

Literature Review

To identify approaches that might be adaptable to our existing ETL design and fulfill the 3 requirements in the previous section, we conducted a first literature review on July 14, 2021; a second one on November 28, 2022; and a third one on February 22, 2023 (Multimedia Appendices 1, 2, and 3). Table 1 includes the search strings and the number of results for three literature databases.

Table . Literature review: database, search string, and number of results.

Database	Search string	Results, n
PubMed	All fields: (incremental) AND ((etl) OR (extract transform load))	7
IEEE Xplore	(("All Metadata": incremental) AND ("All Metadata": etl OR "All Metadata": extract transform load))	15
Web of Science	ALL=(incremental) AND (ALL=(etl) OR ALL=(extract transform load))	46

We included only articles from 2011 to 2022 in English. After removing duplicates, 51 items were left. These were screened independently by two authors (EH and MZ). Through the title and abstract screening, we identified 12 relevant articles. After the screening of the full texts, we included 8 articles within our research. Reasons for excluding the other articles were other meanings of the abbreviation "ETL," ETL tools without regard to theoretical approaches of incremental loads, focus on application instead of ETL process and theoretical approach, and quality and error handling without focus on a theoretical approach.

Only 2 of the 8 articles addressed ETL processes for loading patient data into OMOP CDM. Lynch et al [16] introduced an approach for incremental transformation from the data warehouse to OMOP CDM to prevent incremental load errors. They suggest basing the development on a quality assurance process regarding the data quality framework by Kahn et al [17]. Furthermore, they generated ETL batch tracking ids for each record of data during the transformation to OMOP CDM. For 1:1 mappings, they created custom columns in the standardized OMOP CDM tables, and for 1:n or n:1 mappings, they used a parallel mapping table to store the ETL batch id and a link to the corresponding record in OMOP CDM. Lenert et al [18] describe an automated transformation of clinical data into two CDMs (OMOP and PCORnet database) by using FHIR. Therefore, they use the so-called subscriptions of FHIR resources. These subscriptions trigger a function to create a copy of the FHIR resource and its transmission into another system whenever an FHIR resource is created or updated.

Despite OMOP CDM being the target database, the literature search revealed different concepts for incremental ETL itself. Kathiravelu et al [19] described the caching of new or updated data in a temporary table. Of the 8 articles, 7 described various methods for incremental updates, particularly focusing on change data capture (CDC). All describe different categories of CDC, like timestamp-based, audit column–based, trigger-based, log-based, application programming interface–based, and data-based snapshots [16,18,20-24]: (1) Lynch et al [16] and (2) Lenert et al [18] focused on triggers; (3) Wen [20] focused on timestamps and triggers; (4) Thulasiram and Ramaiah [21] and (5) Sun [22] focused on timestamps; (6) Hu and Dessloch [23] focused on timestamps, audit columns, logs, triggers, and snapshots; and (7) Wei Du and Zou [24] focused on snapshots and MapReduce.

In summary, the literature review revealed adaptable approaches, which can be applied for the implementation of requirements B and C. However, no approaches could be found in the literature for requirement A. For this reason, we have to define a new method to enable both bulk and incremental loading in one ETL process. The concrete integration of the approaches into our existing ETL design is described in more detail in the following sections.

Incremental ETL Process Design

Enabling Both Bulk and Incremental Loading

For the specification, if the ETL process should be executed as bulk or incremental load, we added a new Boolean parameter in the configuration file of the ETL process called APP_BULKLOAD_ENABLED. According to the desired loading option, the parameter has to be adjusted before executing the ETL process, with "true" results in a bulk load and "false" results in an incremental load. During the execution of the ETL process, this parameter is further taken into account for the Reader and Processor of the ETL process [2] to distinguish between the needs of bulk and incremental load (eg, to ensure that the OMOP CDM database is not emptied at the beginning of the ETL process execution during an incremental load).

Focusing on CUD-FHIR Resources Since the Last ETL Execution

Our purpose of incremental loading was to focus only on CUD-FHIR resources since the last time the ETL process was executed (whether as bulk or incremental load). Consequently, the ETL process for incremental load has to filter only CUD-FHIR resources from the source. The literature research showed that there are various CDC approaches to detect changes in the source. In our case, FHIR resources in the FHIR Gateway and FHIR Server contain metadata, such as a timestamp indicating when an FHIR resource was created, updated, or deleted in the source (FHIR Gateway: column last_updated_at; FHIR Server: meta.lastUpdated). That is why we used the timestamp-based CDC approach to filter FHIR resources, which have a timestamp specification after the last ETL execution time.

To ensure the filtering for the incremental load, we added two new parameters in the configuration file of the ETL process: DATA_BEGINDATE and DATA_ENDDATE. Both parameters have to be adjusted before executing the ETL process as incremental load. During the execution, the ETL process takes these two parameters into account and only reads FHIR resources from the source that has a metadata timestamp specification that is in [DATA_BEGINDATE, DATA_ENDDATE].

Guarantee Data Correctness in OMOP CDM

To avoid duplicates in OMOP CDM when processing updated and deleted FHIR resources, their existence in OMOP CDM has to be checked during the processing. The FHIR resources themselves do not have a flag that indicates whether they are new or have been changed. Only deleted FHIR resources can be identified by a specific flag in the metadata. To assess the existence of FHIR resources in OMOP CDM, a comparison of the data of the read FHIR resources with the data already available in OMOP CDM has to be done.

The literature research showed an approach to generate a unique tracking id per source data during the transformation process and its storage in OMOP CDM [16]. We decided against the approach of generating an additional id because FHIR resources already contain two identifying FHIR elements themselves: id and identifier. The id represents the logical id of the resource per resource type while the identifier specifies an identifier that is part of the source data. Both FHIR elements allow the unique identification of an FHIR resource per resource type. However, the standardized OMOP CDM tables do not provide the possibility to store this information from FHIR. Furthermore,

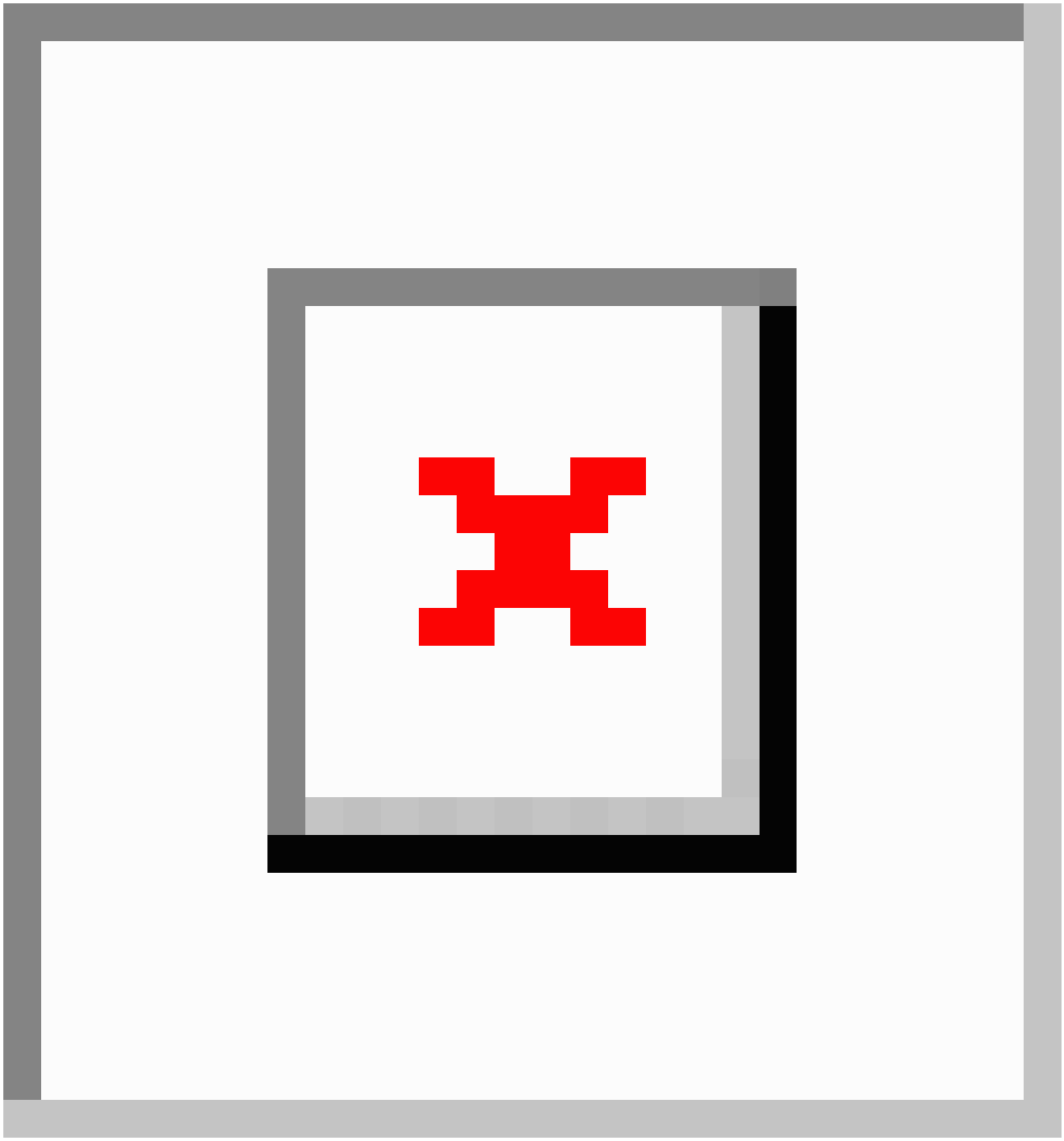
OMOP CDM has its own primary keys for each record in a table independent of the id and identifier used in FHIR. Consequently, after transforming FHIR resources to OMOP CDM, the identifying data from FHIR resources will be lost.

To solve this problem, we need to store the mapping between the id and identifier used in FHIR with the id used in OMOP CDM. Due to the fact that the id of an FHIR resource is only unique per resource type and one FHIR resource can be stored in OMOP CDM in multiple tables, we additionally have to specify the resource type. As mentioned above, Lynch et al [16] presented an approach to store the mapping between tracking ids for source records and ids used in OMOP CDM by using a mapping table and custom columns in OMOP CDM. We have slightly customized this approach and adapted it into our ETL design. Contrary to the use of both mapping tables and custom columns, we considered each approach separately.

Our first approach uses mapping tables for each FHIR resource type in a separate schema in OMOP CDM. With this approach, the Writer of the ETL process has to fill additional mapping tables beside the standardized tables in OMOP CDM. Our second approach focuses on two new columns in the standardized tables in OMOP CDM called “fhir_logical_id” and “fhir_identifier.” These columns store the id and identifier of the FHIR resource. Furthermore, we appended an abbreviation of the resource type as a prefix to the id and identifier of FHIR (eg, “med-” for Medication, “mea-” for MedicationAdministration, or “mes-” for MedicationStatement FHIR resources). In consequence, the combination of the prefix with the id and identifier and its storage in OMOP CDM enables the unique identification of FHIR resources in OMOP CDM. Since the mapping tables and two new columns are required exclusively for the ETL process, the analysis of data across multiple OMOP CDM databases is not affected.

Based on the unique identification of FHIR resources in OMOP CDM, it is now possible to guarantee data correctness in OMOP CDM during incremental loading. Figure 1 shows the exemplary data flow for Condition FHIR resources for the second approach with two new columns. First, the Processor extracts the id and identifier used in FHIR. After that, the prefix is added to both values. Regardless of whether the data was created, updated, or deleted in the source, the ETL process next verifies each processed FHIR resource’s existence in OMOP CDM using the mapping tables or two new columns. During the verification, records are deleted in OMOP CDM if they were found. This approach is also used for updated FHIR resources to avoid incomplete updates for cross-domain mappings in OMOP CDM. Consequently, we do not perform updates on the existing records in OMOP CDM except Patient and Encounter FHIR resources to ensure referential integrity in OMOP CDM. In case FHIR resources are marked as deleted in the source, the processing is completed. Otherwise, the same semantic mapping logic as for bulk loading [2] applies afterward, and the data of the FHIR resources are written to OMOP CDM as new records with new OMOP ids.

Figure 1. Excerpt of the data flow of the Condition Processor. CDM: Common Data Model; FHIR: Fast Healthcare Interoperability Resources; OMOP: Observational Medical Outcomes Partnership.



Evaluation of the Incremental Load Process

For the evaluation of the incremental load process, we defined and executed two ETL test designs. First, we tested which approach to store the mapping between id and identifier used in FHIR with the id used in OMOP CDM was the most performant. For this purpose, we implemented a separate ETL process version for each approach. Afterward, we executed the ETL process as bulk load first and as incremental load afterward, and compared the execution times between the mapping table approach and the column approach. For further evaluation of the incremental load process, we have chosen the most performant approach, resulting in a new optimized ETL process version for the second ETL test design.

To test the achievement of the 3 requirements identified during the initial analysis of our ETL process, we defined and executed a second ETL test design (Table 2) that compares the results of bulk loading with those of incremental loading regarding performance and data correctness. Our hypotheses here are that the execution time of incremental loading alone is less than bulk loading including daily updates and that the amount of data per table in OMOP CDM is identical after incremental loading and bulk loading, including daily updates.

Table . Extract-Transform-Load test design regarding performance and data correctness.

Test focus	Hypothesis
Performance	t(bulk loading (3 mon)) + t(incremental loading (1 d)) < t(bulk loading (3 mon)) + t(bulk loading (3 mon + 1 d))
Data correctness	#((bulk loading (3 mon)) + (incremental loading (1 d))) = #(bulk loading (3 mon + 1 d))

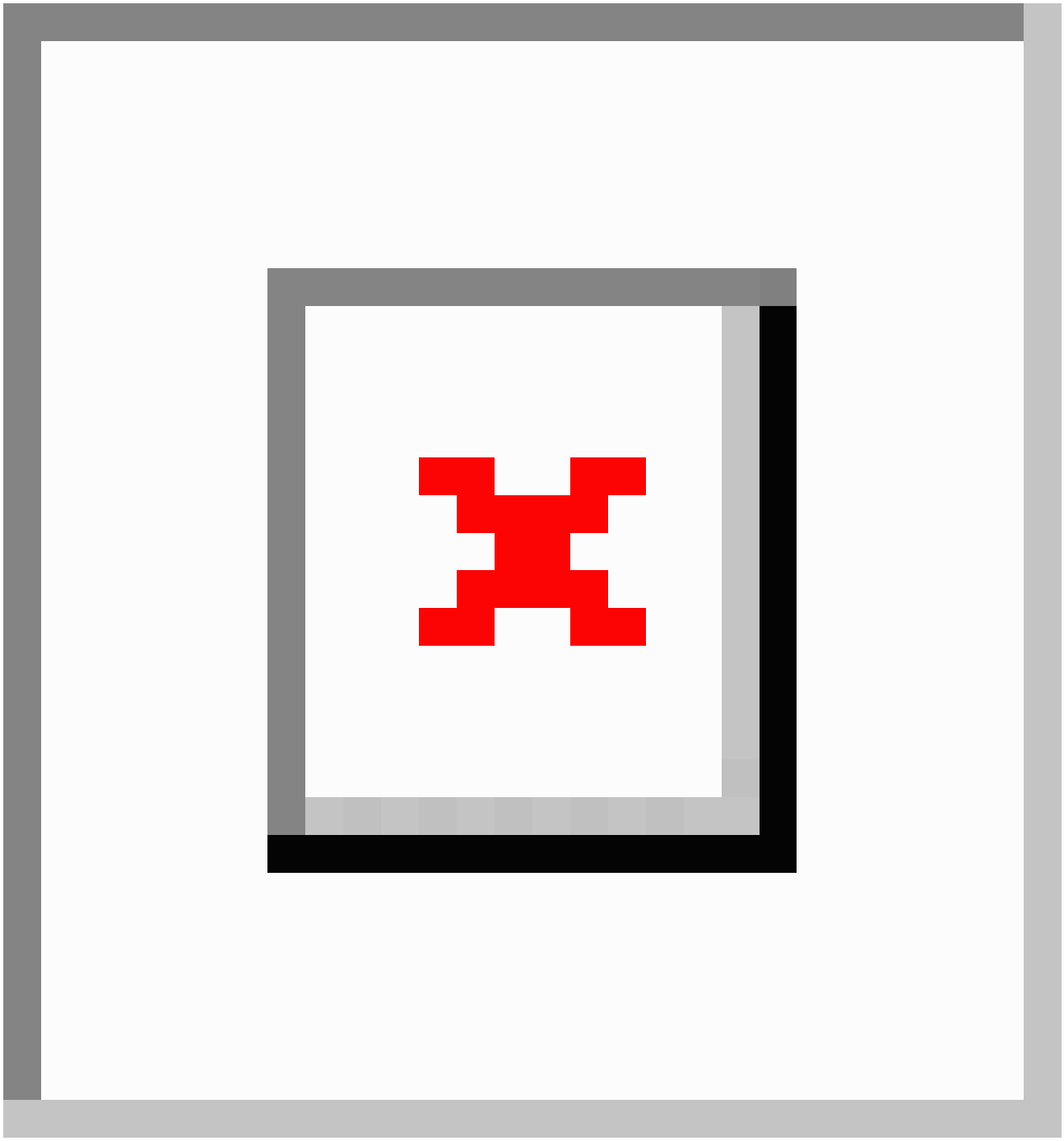
For both ETL test designs, we used a total of 3,802,121 synthetic FHIR resources version R4 based on the MI-I CDS version 1.0, which were generated using random values. Furthermore, we simulated CUD-FHIR resources for testing incremental loading for 1 day. For the simulation, we checked the frequency distribution of CUD data per domain in our source system with real-world data for 8 days and calculated the average value (see Multimedia Appendix 4). In addition, we set up one OMOP CDM v5.3.1 database as the target and executed the ETL process according to our test designs. For both ETL tests, we tracked the execution times based on the time stamps in the logging file of the ETL process until the corresponding job finished successfully. In a second step, we recorded the data quantity for each filled table in OMOP CDM and compared the results between the two ETL loading options for the second ETL test.

Results

Architecture of the ETL Process

The implemented ETL process extension for incremental loading of FHIR resources to OMOP CDM has not changed the basic architecture of the ETL process as proposed by Peng et al [2], consisting of Reader, Processor, and Writer (Figure 2). The only addition is a switch at the beginning of the ETL process, which allows the user to select between bulk load and incremental load (requirement A). Moreover, we configured the Reader for incremental loading of CUD-FHIR resources on a daily basis (requirement B). In the Processor, we added the logic of the verification of CUD-FHIR resources and their deletion from OMOP CDM if they already exist (requirement C).

Figure 2. Architecture of the FHIR-to-OMOP Extract-Transform-Load process including incremental load. CDM: Common Data Model; FHIR: Fast Healthcare Interoperability Resources; OMOP: Observational Medical Outcomes Partnership.



The ETL process covering bulk and incremental load is available in the OHDSI repository ETL-German-FHIR-Core [25].

Findings of the First ETL Test

The first ETL test focused on the performance measurement of the mapping table approach versus the column approach. First, we executed both ETL approaches as a bulk load. The column approach took about 30 minutes to transform FHIR resources to OMOP CDM. In contrast, the mapping table approach was still not finished after 4 hours. Therefore, we stopped the ETL execution and did not test the incremental loading anymore. Consequently, for the incremental ETL design, we decided to use the column approach due to its better performance and executed the subsequent performance evaluations with it.

Findings of the Second ETL Test

The second ETL test dealt with testing our two hypotheses in Table 2. First, we compared the execution times between a bulk load (3 mon + 1 d) and an initial bulk load (3 mon) followed by an incremental load (1 d). For this, each loading option was executed three times. Based on the results, we calculated the average execution times. The performance results (Multimedia Appendix 5) show that an initial bulk load (13.31 min) followed by a daily incremental load (2.12 min) is more efficient than an everyday full load (17.07 min). Looking at the percentage improvement in performance, it can be shown that incremental loading had 87.5% less execution time than a daily full load

(2.12 min compared to 17.07 min). Referring to our first hypothesis, we were able to prove our initial assumption.

After the execution of both loading options, we further checked the data quantity for each filled table in OMOP CDM and compared the results of it. As shown in Table 3, both loading options resulted in the same amount of data (Multimedia Appendix 5). Consequently, we were also able to confirm our second hypothesis regarding data correctness in OMOP CDM.

Table . Results of the data quantity comparison in the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) between bulk and incremental load.

	Bulk load (3 mon + 1 d), n	Bulk load (3 mon) + incremental load (1 d), n
Care_site	152	152
Condition_occurrence	800,640	800,640
Death	857	857
Drug_exposure	1,171,521	1,171,521
Fact_relationship	2,323,894	2,323,894
Measurement	231,369	231,369
Observation	511,844	511,844
Observation_period	15,037	15,037
Person	15,037	15,037
Procedure_occurrence	168,384	168,384
Source_to_concept_map	251	251
Visit_detail	43,929	43,929
Visit_occurrence	29,898	29,898

Discussion

Principal Findings

Based on the partial results for research questions 1 and 2, we defined three methods to integrate incremental loading into our ETL process design. In this context, the 3 identified requirements from the initial analysis could be implemented by taking existing approaches from the literature into account (research question 3). Moreover, the incremental load process was tested at 10 university hospitals in Germany and ensures daily data transfer to OMOP CDM for the CTRSS. This proves that the ETL process is also suitable for real-world data, although it was developed with synthetic data.

Currently, our ETL process requires FHIR resources following the MI-I core data set specification. However, our initial requirements analysis showed that the implemented incremental ETL logic does not affect the semantic mapping from FHIR to OMOP CDM described by Peng et al [2]. In consequence, the incremental ETL logic is independent of the data available in the FHIR format. Therefore, it can be used to incrementally transform international FHIR profiles such as the US Core Profiles [26] to OMOP CDM.

Limitations

Nevertheless, the ETL process has limitations in its execution capabilities. As our FHIR resources comprise a logical id corresponding to the id in our source system, our ETL process is currently not able to deal with changing server end points, resulting in changing logical ids. Additionally, so far, we have not included an option to automatically start incremental loading nor do we support real-time streaming (eg, via Apache Kafka [27]). These limitations are part of future work.

To evaluate incremental loading compared to bulk loading, we performed two ETL tests (research question 4). The results of the performance tests showed that the column approach is more performant than the mapping table approach. Our suspected explanation for this is that the mapping table approach requires additional tables to be filled besides the standardized tables in OMOP CDM. Consequently, during the verification of FHIR resources in OMOP CDM, a lookup and deletion in several tables (mapping tables and standardized tables) is necessary, whereas the column approach only accesses the standardized tables.

Referring to our two hypotheses regarding performance and data correctness between incremental loading and bulk loading, we showed that our initial assumptions were proven. With the option of an incremental ETL process, we were able to reduce execution times to provide data in OMOP CDM on a daily basis, without data loss compared to the bulk load ETL process. For our future work, we want to further evaluate at what point bulk load is more worthwhile than incremental loading. The results of these evaluations will be incorporated into the automation concept.

During the productive use of the ETL process, we identified two issues that have to be considered in the context of incremental loading to OMOP CDM. First, OHDSI provides a wide range of open-source tools (eg, ATLAS [28]) for cohort definitions or statistical analyses. To make ATLAS work on the data in OMOP CDM, a summary report has to be generated in advance using ACHILLES [29] (Automated Characterization of Health Information at Large-Scale Longitudinal Evidence

Systems), an R package that provides characterization and visualization. Regarding the incremental ETL process, ACHILLES has to be run after each successful execution of the incremental ETL process.

A second issue that needs to be addressed relates to the ids in the standardized tables in OMOP CDM. The incremental loading process requires the assignment of new ids in the OMOP CDM. While this was not a problem during development, it becomes obvious when a large amount of data is processed. In this context, the maximum id in the tables of OMOP CDM was reached, which led to a failure of the ETL process. We need to pay special attention to this point and find a solution (eg, by reusing deleted ids or by changing the ETL process in a real updating ETL process). As this problem does not occur during the bulk load process, a current workaround is to start that process if the incremental load fails, which is possible as our process comprises bulk and incremental load options.

Conclusions

The presented ETL process from FHIR to OMOP CDM now enables both bulk and incremental loading. To receive daily updated recruitment proposals with the CTRSS, the ETL process no longer needs to be executed as a bulk load every day. One initial load supplemented by incremental loads per day meets the requirements of the CTRSS while being more performant. Moreover, since the incremental ETL logic is not restricted to the MI-I CDS specification, it can also be used for international studies that require daily updated data from FHIR resources in OMOP CDM. To be able to use not only the logic of incremental loading internationally, but the whole ETL process itself, the support of arbitrary FHIR profiles is needed. This requires a modularization and generalization of current ETL processes. For that, we will evaluate the extension to metadata-driven ETL in the near future.

Acknowledgments

The research reported in this work was accomplished as part of the German Federal Ministry of Education and Research within the Medical Informatics Initiative, Medical Informatics in Research and Care in University Medicine (MIRACUM) consortium (FKZ: 01lZZ180L; Dresden). The article processing charge was funded by the joint publication funds of the Technische Universität, Dresden, including the Carl Gustav Carus Faculty of Medicine, and the Sächsische Landesbibliothek – Staats- und Universitätsbibliothek, Dresden, as well as the Open Access Publication Funding of the Deutsche Forschungsgemeinschaft.

Authors' Contributions

All authors contributed substantially to this work. EH and MZ conducted the literature review. EH and YP contributed to the Extract-Transform-Load (ETL) process design and implementation. EH, YP, IR, and FB reviewed the ETL process design and implementation. YP contributed to the ETL process execution and evaluation. EH prepared the original draft. EH, YP, IR, FB, MZ, and MS reviewed and edited the manuscript. MS contributed toward the resources. All authors have read and agreed to the current version of the manuscript and take responsibility for the scientific integrity of the work.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Results and screenings of the literature review on July 14, 2021.
[XLSX File, 42 KB - medinform_v11i1e47310_app1.xlsx]

Multimedia Appendix 2
Results and screenings of the literature review on November 28, 2022.
[XLSX File, 44 KB - medinform_v11i1e47310_app2.xlsx]

Multimedia Appendix 3
Results and screenings of the literature review on February 22, 2023.
[XLSX File, 47 KB - medinform_v11i1e47310_app3.xlsx]

Multimedia Appendix 4
Frequency distribution of created, updated, and deleted data.
[XLSX File, 11 KB - medinform_v11i1e47310_app4.xlsx]

Multimedia Appendix 5
Results of the second Extract-Transform-Load test.
[XLSX File, 24 KB - medinform_v11i1e47310_app5.xlsx]

References

1. Hariton E, Locascio JJ. Randomised controlled trials—the gold standard for effectiveness research. BJOG: Int J Obstet Gy 2018 Dec;125(13):1716-1716. [doi: 10.1111/1471-0528.15199] [Medline: 29916205]
2. Peng Y, Henke E, Reinecke I, Zoch M, Sedlmayr M, Bathelt F. An ETL-process design for data harmonization to participate in international research with German real-world data based on FHIR and OMOP CDM. Int J Med Inform 2023 Jan;169:104925. [doi: 10.1016/j.ijmedinf.2022.104925] [Medline: 36395615]
3. Observational Health Data Sciences and Informatics. Standardized data: the OMOP common data model. URL: www.ohdsi.org/data-standardization/ [accessed 2022-11-7]
4. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. Stud Health Technol Inform 2015;216:574-578. [Medline: 26262116]]
5. Semler SC, Wissing F, Heyder R. German Medical Informatics Initiative. Methods Inf Med 2018 Jul;57(S 01):e50-e56. [doi: 10.3414/ME18-03-0003] [Medline: 30016818]
6. Kabachinski J. What is Health Level 7? Biomed Instrum Technol 2006 ;40(5):375-379. [doi: 10.2345/i0899-8205-40-5-375.1] [Medline: 17078369]
7. Medical Informatics Initiative. The Medical Informatics Initiative's core data set. URL: www.medizininformatik-initiative.de/en/medical-informatics-initiatives-core-data-set [accessed 2022-11-7]
8. GitHub. FHIR gateway. 2023. URL: github.com/miracum/fhir-gateway [accessed 2023-03-15]
9. GitHub. HAPI FHIR. 2023. URL: github.com/hapifhir/hapi-fhir [accessed 2023-03-15]
10. GitHub. Blaze. 2023. URL: github.com/samply/blaze [accessed 2023-03-15]
11. OMOP Common Data Model. OMOP CDM V5.3.1. URL: ohdsi.github.io/CommonDataModel/cdm531.html [accessed 2022-11-7]
12. Spring. Spring Batch - reference documentation. URL: docs.spring.io/spring-batch/docs/current/reference/html/index.html [accessed 2022-11-7]
13. Observational Health Data Sciences and Informatics. The Book of OHDSI. 2021. URL: ohdsi.github.io/TheBookOfOhdsi/ [accessed 2022-04-19]
14. Prokosch HU, Acker T, Bernarding J, Binder H, Boeker M, Boerries M, et al. MIRACUM: Medical Informatics in Research and Care in University Medicine. Methods Inf Med 2018 Jul;57(S 01):e82-e91. [doi: 10.3414/ME17-02-0025] [Medline: 30016814]
15. Reinecke I, Gulden C, Kümmel M, Nassirian A, Blasini R, Sedlmayr M. Design for a modular clinical trial recruitment support system based on FHIR and OMOP. Stud Health Technol Inform 2020 Jun 16;270:158-162. [doi: 10.3233/SHTI200142] [Medline: 32570366]
16. Lynch KE, Deppen SA, DuVall SL, Viernes B, Cao A, Park D, et al. Incrementally transforming electronic medical records into the observational medical outcomes partnership common data model: a multidimensional quality assurance approach. Appl Clin Inform 2019 Oct;10(5):794-803. [doi: 10.1055/s-0039-1697598] [Medline: 31645076]
17. Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, et al. A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. EGEMS (Wash DC) 2016 Sep 11;4(1):1244. [doi: 10.13063/2327-9214.1244] [Medline: 27713905]
18. Lenert LA, Ilatovskiy AV, Agnew J, Rudisill P, Jacobs J, Weatherston D, et al. Automated production of research data marts from a canonical Fast Healthcare Interoperability Resource data repository: applications to COVID-19 research. J Am Med Inform Assoc 2021 Jul 30;28(8):1605-1611. [doi: 10.1093/jamia/ocab108] [Medline: 33993254]
19. Kathiravelu P, Sharma A, Galhardas H, Roy P, Veiga L. On-demand big data integration: a hybrid ETL approach for reproducible scientific research. Distributed Parallel Databases 2018 Apr 24(37):273-295. [doi: 10.1007/s10619-018-7248-y]
20. Wen WJ. Research on the incremental updating mechanism of marine environmental data warehouse. Appl Mechanics Materials 2014 Oct;668-669:1378-1381. [doi: 10.4028/www.scientific.net/AMM.668-669.1378]
21. Thulasiram S, Ramaiah N. Real time data warehouse updates through extraction-transformation-loading process using change data capture method. 2019 Presented at: Second International Conference on Computer Networks and Communication Technologies; May 23-24; Coimbatore, India p. 552-560. [doi: 10.1007/978-3-030-37051-0]
22. Sun YY. Research and implementation of an efficient incremental synchronization method based on Timestamp. 2022 Presented at: 3rd International Conference on Computing, Networks and Internet of Things; May 20-22; Qingdao, China p. 158-162. [doi: 10.1109/CNIOT55862.2022.00035]
23. Hu Y, Dessloch S. Extracting deltas from column oriented NoSQL databases for different incremental applications and diverse data targets. Data Knowledge Eng 2014 Sep;93:42-59. [doi: 10.1016/j.datak.2014.07.002]
24. Wei Du D, Zou X. Differential snapshot algorithms based on hadoop mapreduce. 2015 Presented at: 12th International Conference on Fuzzy Systems and Knowledge Discovery; August 15-17; Zhangjiajie, China p. 1203-1208. [doi: 10.1109/FSKD.2015.7382113]
25. GitHub. FHIR-to-OMOP. 2023. URL: github.com/OHDSI/ETL-German-FHIR-Core [accessed 2023-03-15]
26. Health Level Seven International. US Core Implementation Guide. URL: www.hl7.org/fhir/us/core/ [accessed 2023-03-15]
27. Apache Kafka. URL: kafka.apache.org/ [accessed 2023-03-15]
28. GitHub. ATLAS. 2023. URL: github.com/OHDSI/Atlas [accessed 2023-03-15]

29. GitHub. Achilles. 2023. URL: github.com/OHDSI/Achilles [accessed 2023-03-15]

Abbreviations

ACHILLES: Automated Characterization of Health Information at Large-Scale Longitudinal Evidence Systems
CDC: change data capture
CDM: Common Data Model
CDS: Core Data Set
CTRSS: clinical trial recruitment support system
CUD: created, updated, or deleted
ETL: Extract-Transform-Load
FHIR: Fast Healthcare Interoperability Resources
HL7: Health Level Seven
MI-I: Medical Informatics Initiative Germany
MIRACUM: Medical Informatics in Research and Care in University Medicine
OHDSI: Observational Health Data Sciences and Informatics
OMOP: Observational Medical Outcomes Partnership

Edited by C Lovis; submitted 15.03.23; peer-reviewed by A Kiourtis, F Amar; revised version received 04.04.23; accepted 03.05.23; published 21.08.23.

Please cite as:
Henke E, Peng Y, Reinecke I, Zoch M, Sedlmayr M, Bathelt F
An Extract-Transform-Load Process Design for the Incremental Loading of German Real-World Data Based on FHIR and OMOP CDM: Algorithm Development and Validation
JMIR Med Inform 2023;11:e47310
URL: https://medinform.jmir.org/2023/1/e47310
doi:10.2196/47310

Interoperable, Domain-Specific Extensions for the German Corona Consensus (GECCO) COVID-19 Research Data Set Using an Interdisciplinary, Consensus-Based Workflow: Data Set Development Study

Gregor Lichtner[1,2,3], PhD; Thomas Haese[1], MSc; Sally Brose[1], MSc; Larissa Röhrig[1,4], MSc; Liudmila Lysyakova[5,6], PhD; Stefanie Rudolph[5,6], PhD; Maria Uebe[5,6], BA; Julian Sass[1], MSc; Alexander Bartschke[1], MSc; David Hillus[7], MD; Florian Kurth[7,8,9], MD; Leif Erik Sander[7], MD; Falk Eckart[10], MD; Nicole Toepfner[10], MD; Reinhard Berner[10], MD; Anna Frey[11], MD; Marcus Dörr[12], MD; Jörg Janne Vehreschild[13,14,15], MD; Christof von Kalle[5,6], MD; Sylvia Thun[1], MD

Corresponding Author:
Gregor Lichtner, PhD

Abstract

Background: The COVID-19 pandemic has spurred large-scale, interinstitutional research efforts. To enable these efforts, researchers must agree on data set definitions that not only cover all elements relevant to the respective medical specialty but also are syntactically and semantically interoperable. Therefore, the German Corona Consensus (GECCO) data set was developed as a harmonized, interoperable collection of the most relevant data elements for COVID-19–related patient research. As the GECCO data set is a compact core data set comprising data across all medical fields, the focused research within particular medical domains demands the definition of extension modules that include data elements that are the most relevant to the research performed in those individual medical specialties.

Objective: We aimed to (1) specify a workflow for the development of interoperable data set definitions that involves close collaboration between medical experts and information scientists and (2) apply the workflow to develop data set definitions that include data elements that are the most relevant to COVID-19–related patient research regarding immunization, pediatrics, and cardiology.

Methods: We developed a workflow to create data set definitions that were (1) content-wise as relevant as possible to a specific field of study and (2) universally usable across computer systems, institutions, and countries (ie, interoperable). We then gathered medical experts from 3 specialties—infectious diseases (with a focus on immunization), pediatrics, and cardiology—to select data elements that were the most relevant to COVID-19–related patient research in the respective specialty. We mapped the data elements to international standardized vocabularies and created data exchange specifications, using Health Level Seven International (HL7) Fast Healthcare Interoperability Resources (FHIR). All steps were performed in close interdisciplinary collaboration with medical domain experts and medical information specialists. Profiles and vocabulary mappings were syntactically and semantically validated in a 2-stage process.

Results: We created GECCO extension modules for the immunization, pediatrics, and cardiology domains according to pandemic-related requests. The data elements included in each module were selected, according to the developed consensus-based workflow, by medical experts from these specialties to ensure that the contents aligned with their research needs. We defined data set specifications for 48 immunization, 150 pediatrics, and 52 cardiology data elements that complement the GECCO core data set. We created and published implementation guides, example implementations, and data set annotations for each extension module.

Conclusions: The GECCO extension modules, which contain data elements that are the most relevant to COVID-19–related patient research on infectious diseases (with a focus on immunization), pediatrics, and cardiology, were defined in an interdisciplinary, iterative, consensus-based workflow that may serve as a blueprint for developing further data set definitions. The GECCO extension modules provide standardized and harmonized definitions of specialty-related data sets that can help enable interinstitutional and cross-country COVID-19 research in these specialties.

(JMIR Med Inform 2023;11:e45496) doi:10.2196/45496

KEYWORDS

interoperability; research data set; Fast Healthcare Interoperability Resources; FHIR; FAIR principle; COVID-19; interoperable; SARS-CoV-2; pediatric; immunization; cardiology; standard

Introduction

The COVID-19 pandemic has led to unprecedented, strong efforts in connecting nationwide and international research to help manage the disease and its effects on public health. To enable research across different health care providers, institutions, or even countries, interoperability between medical data systems is essential [1]. Therefore, early in the pandemic, the German Corona Consensus (GECCO) data set was developed in a collaborative effort to provide a standardized, unified core data set for interinstitutional COVID-19–related patient research [2]. The GECCO data set specifies a set of 81 essential clinical data elements from 13 domains, such as anamnesis and risk factors, symptoms, and vital signs, that have been selected by expert committees from university hospitals, professional associations, and research initiatives. Since its development, the GECCO data set has been implemented in a large number of institutions, most notably in virtually all German university hospitals, which now provide access to the GECCO data set in the context of the German COVID-19 Research Network of University Medicine ("Netzwerk Universitätsmedizin") [3,4].

The GECCO data set was developed to contain as many relevant data elements as possible but few enough to keep the effort of implementing the data set manageable. Therefore, the data set contains mostly data elements of general research interest, excluding data elements that are only of interest for particular medical specialties or use cases. These data items are considered part of domain-specific extension modules of the GECCO data set, which are introduced in this paper.

We aimed to develop domain-specific extensions to the GECCO data set that cover the most relevant data elements for COVID-19–related patient research in the infectious disease (with a focus on immunization), pediatrics, and cardiology medical specialties. To that end, we first developed a workflow that aims at providing data set definitions that (1) contain the most relevant data elements for the research aims of the end users and (2) can be applied universally across institutions and countries. We then followed that workflow with different groups of medical experts from different medical specialties to define extension modules that are relevant for research regarding immunization, pediatrics, and cardiology.

These extension modules complement the GECCO core data set and use the same international health IT standards and terminologies as those in the GECCO data set, such as the Systematized Nomenclature of Medicine-Clinical Terms (SNOMED CT) [5], the Logical Observation Identifiers Names and Codes (LOINC) [6,7], and the Fast Healthcare Interoperability Resources (FHIR) [8,9] standard. The extension modules were developed in close alignment with the GECCO data set to ensure interoperability and compatibility with existing definitions.

We herein describe the consensus-based data element selection and data format definition workflow that we applied in close collaboration with medical experts from 3 specialties—infectious diseases (with a focus on immunization), pediatrics, and cardiology (ie, for content definition)—as well as medical information specialists and FHIR developers (ie, for technical aspects). This workflow may serve as a blueprint for the further development of consensus-based data set definitions.

Methods

Workflow Definition

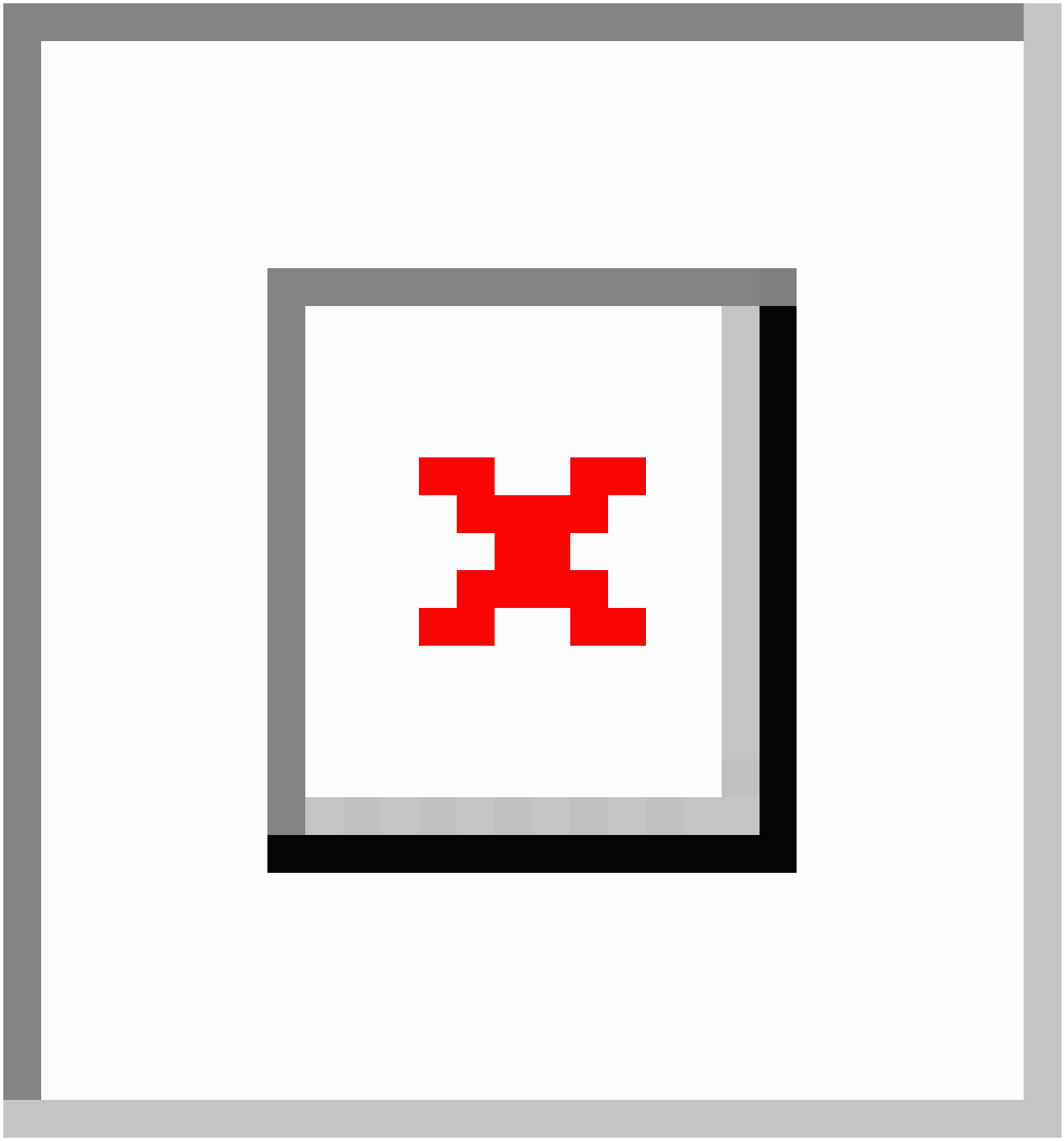
We aimed to develop a workflow to create data set definitions that are (1) content-wise as relevant as possible to a specific field of study and (2) universally usable across computer systems, institutions, and countries (ie, interoperable). We based the specification of the workflow on our experience with the definition of the GECCO data set, during which health professionals from 50 institutions (university hospitals, professional associations, and other relevant organizations) participated to define the most relevant data elements for general-scope, COVID-19–related research [2]. To fulfill the first requirement (relevancy), we decided to leave the full responsibility of data element selection to groups of medical professionals of the respective specialty, with only minimal interference by the medical information specialists. We deliberately did not specify the exact process of how the group of medical experts could select the data elements (eg, literature

review, focus groups, and consensus-based processes) to allow for the maximal flexibility of the data set definition workflow, with respect to the medical experts' values and preferences. To fulfill the second requirement (interoperability), we adopted a model that was loosely based on the data FAIRification workflow of Jacobsen et al [10]; the mapping, quality assurance, and publication steps are outlined in detail below.

Selection of Data Items

The content of the domain-specific research data sets was defined by medical domain experts in a transparent workflow (Figure 1). The involvement of the medical domain experts as the end users of the data to be provided ensured that the contents of the data sets were aligned to the actual research needs. In our project, the so-called *subject- and organ-specific working groups* of the National Pandemic Cohort Network ("Nationales Pandemie Kohorten Netz" [NAPKON]) served as the domain-specific groups of medical experts. These groups were established by a voluntary association of medical experts from the medical specialties within the nationwide NAPKON project in Germany. Each of the subject- and organ-specific working groups elected a board, and all communication between the data set developers and the working groups was organized and carried out via the working groups' boards. In preparation for the GECCO extension modules, we invited the subject- and organ-specific groups for infectious diseases (with a focus on immunization), pediatrics, and cardiology to provide up to 50 data elements (with up to 10 response items each) that were, in the view of the medical experts, the most relevant to patient-related COVID-19 research in these medical specialties and not already included in the GECCO core data set. If necessary, more data items or response options could be provided in coordination with the medical information specialists. The provided data items were then reviewed by the medical information specialists, and a first definition of the contents of the extension module was returned to the respective subject- and organ-specific working group for approval or change requests. After approval by the subject- and organ-specific working group, the definition of the extension module content was considered finalized.

Figure 1. Flowchart of the consensus-based, interdisciplinary data set definition and mapping workflow for the domain-specific COVID-19 research data sets. FHIR: Fast Healthcare Interoperability Resources.



Development of the Standardized Data Formats

To map the data items selected by the subject- and organ-specific working groups to international standard vocabularies, we performed a consensus-based mapping procedure, wherein every concept was mapped to appropriate vocabularies—the SNOMED CT for general concepts [11]; LOINC for observations [7]; *International Statistical Classification of Diseases and Related Health Problems, 10th Revision, German Modification* for diagnoses [12]; Anatomical Therapeutic Chemical Classification System for Germany for drugs and active ingredients [13]; and Unified Code for Units of Measure for measurement units [14]—by 2 medical information specialists independently. Ambiguities and nonmatching mappings were then discussed among the medical information specialists and in close collaboration with the medical experts of the subject- and organ-specific working groups until consensus was achieved. The data item–to-concept mappings were annotated on ART-DECOR, an open-source collaboration platform for creating and maintaining data set element descriptions [15].

As for the GECCO data set, the format for data exchange was specified by using Health Level Seven International (HL7) FHIR resources. The mapping of data items to FHIR resources was performed in an iterative, consensus-based workflow among the medical information specialists. Wherever possible, published FHIR profiles from the GECCO data set, the Medical

Informatics Initiative [16], or the National Association of Statutory Health Insurance Physicians ("Kassenärztliche Bundesvereinigung") [17]—in this order of priority—served as the base definition for the future extension module profiles.

The profiles and value sets were specified by using the FHIR Shorthand (FSH) language (version 1.2.0) and translated to Structure Definition JSON files by using the HL7 FSH SUSHI software package (version 2.2.3) [18,19]. We required that at least one exemplary instance be defined for every profile. The syntactic validation of the profile and value set definitions was performed through the error-free conversion of the FSH files to JSON via SUSHI, and the subsequent validation of each profile and their defined instances was performed by using the HL7 FHIR validator as implemented in the FSH Validator Python package (version 0.2.2) [20]. After the successful syntactic validation of a set of profiles, the profiles were subjected to a 2-stage review process, as follows. First, the profiles and the corresponding value sets and extensions were internally reviewed for semantic appropriateness with the GECCO core developer (JS). After all necessary changes and approval by the internal reviewer, the profiles were subjected to the second review round by an external FHIR development expert. Subsequent to necessary corrections and the approval of the external reviewer, the respective profiles, together with their value sets and, optionally, extensions and code systems, were considered finalized and published to the main branch of the Git repository. The subsequent and ongoing maintenance phase of the data set definitions involves inviting implementers and users to report any issues that they encounter with the definitions, in order to ensure their accuracy and relevance over time.

The whole development process was performed collaboratively on GitHub. The syntactic validation of the profiles was performed via continuous integration/continuous development workflows, which were implemented as GitHub actions. Semantic validation during the internal and external review rounds was performed by using pull requests to 2 different Git branches. After the final approval, profiles and value sets were merged into the main branch of the respective extension module's repository, which served as the publication branch of that module. Since then, maintenance requests and updates of the extension modules have been handled via GitHub issues. All kinds of relevant changes have become subjects of the internal review, as defined above; major changes (eg, nontechnical corrections) are additionally exposed to the external review.

Implementation guides were created for all 3 extension modules, using the FHIR IG Publisher tool and a customized template for the implementation guides' HTML pages [21]. The implementation guides were published to GitHub pages, where they remain automatically synchronized with the main branch of the respective repository via continuous integration/continuous development workflows.

Ethics Approval

This study did not involve any human or animal experiments. No permissions were required to access any data used in this study.

Results

Data Set Definition Workflow

We developed an interdisciplinary, iterative, expert consensus–based workflow for the initial definition of domain-specific COVID-19 research data sets based on 2 key requirements. The first key requirement for the content of the data sets was that the content definition (ie, selection of data elements) was to be performed under the full responsibility of a group of medical experts to ensure that the selected data elements were truly those that are required for research in the respective medical specialty. The second key requirement was to produce FAIR (Findable, Accessible, Interoperable, Reusable) digital assets [22], that is, the data set definitions should be represented in FHIR profiles and implementation guides, and these should be registered on open platforms (ie, findable); they should be retrievable through open, free, standard protocols (ie, accessible); they should use only standard, international medical terminologies, such as SNOMED CT and LOINC (ie, interoperable); and they should be released with rich usage guides and examples (FHIR implementation guide) and under a permissive license (ie, reusable).

To fulfill these requirements, the data set definition workflow consists of the following 6 phases: content definition, mapping, quality assurance, publication, an optional public review, and maintenance (Figure 1). In the content definition phase, a group of medical experts from a particular medical specialty are approached by the medical information specialists and asked to provide a list of the data elements that are the most relevant to patient-related COVID-19 research in the respective medical specialty. How the medical expert group compiles the list in detail is left to their discretion (eg, based on systematic literature review or Delphi consensus processes). The medical information specialists only review the provided lists for consistency and redundancy and compile the final content definition in agreement with the medical expert group. In the mapping phase, all data elements are then mapped to international terminologies in consultation with the group of medical experts. Based on these, a logical model and the mappings of data elements to FHIR resources are established. In the quality assurance phase, the FHIR specifications are syntactically validated by using the HL7 FHIR validator as implemented in the FSH Validator Python package (version 0.2.2) [20] and then subjected to a 2-stage review process, during which 2 individual data interoperability and harmonization experts validate the specifications semantically, that is, they validate that the data elements defined by the group of medical experts are appropriately mapped to international standards. After any required changes, the logical model and the FHIR implementation guide are published and are openly accessible to the research community in repositories that fulfill the FAIR criteria as closely as possible, such as ART-DECOR [15] for the logical model and GitHub or the FHIR Implementation Guide registry for the implementation guide [23]. If desired, the initial release of the data set definition can be subjected to public review and balloting processes, which allow stakeholders to provide feedback and suggest changes. The public review and balloting processes provide an opportunity to obtain broader

input from and facilitate consensus building among the research community and stakeholders. Any changes resulting from the review and balloting processes can then be incorporated into the data set definition according to the herein presented workflow, and the updated version is released and maintained according to the same workflow. In the maintenance phase, the medical information specialists invite implementers and users of the data set definitions to report any issues they encounter with the definitions via GitHub issues or email, in order to ensure their accuracy and relevance over time. During the maintenance phase, requests for changes or updates to the data set definition should generally be limited to minor issues or corrections, as adding new data elements or making significant modifications to the definition would require running the entire workflow from the beginning.

Data Set Contents

Groups of Medical Experts

In the context of the NAPKON project of the German COVID-19 Research Network of University Medicine [24], so-called *subject- and organ-specific working groups* were established by the voluntary association of medical experts from different medical specialties. In preparation for the domain-specific data set definitions that extend the GECCO core data set, the working groups for infectious diseases (with a focus on immunization), pediatrics, and cardiology were invited by the data set development group to provide up to 50 data elements (with up to 10 response items each) that were of particular interest to their field, concerned patient-related COVID-19 research, and were not already included in the GECCO core data set. For the immunization data set definition, physicians from the COVIM (Collaborative Immunity Platform of the Netzwerk Universitätsmedizin) study for the determination and use of SARS-CoV-2 immunity [25-27] assumed the role of the organ-specific working group, as no such working group had been established previously.

Overview

We extended the GECCO core data set by developing domain-specific data set definitions for a total of 250 new data items—48 for the immunization extension module, 150 for the pediatrics extension module, and 52 for the cardiology extension module. These data items were collected, via an iterative consensus-based approach, from the subject- and organ-specific working groups, and they fall under 10 of the 13 data categories of the GECCO data set (Table 1). Data elements and the number of items for each individual extension module are shown in Tables 2, 3, and 4. The full lists of items are shown in the Tables S1-S3 in Multimedia Appendix 1.

Table . Number of data items per GECCO[a] data set category for each extension module.

GECCO data category	GECCO extension module		
	Immunization data items (N=48), n	Pediatrics data items (N=150), n	Cardiology data items (N=52), n
Anamnesis and risk factors	13	21	6
Complications	24	47	7
Demographics	__[b]	6	—
Epidemiological factors	—	—	—
Imaging	—	2	36
Laboratory values	1	27	2
Medication	1	35	1
Onset of illness and admission	6	2	—
Outcome at discharge	—	—	—
Study enrollment and inclusion criteria	—	—	—
Symptoms	—	9	—
Therapy	2	1	—
Vital signs	1	—	—

[a]GECCO: German Corona Consensus.

[b]Not available.

Table . Types of data elements in the immunization extension module extending the GECCO[a] core data set. Shown are the data elements and the FHIR[b] resource they have been mapped to, as well as the number of items for each data element (ie, different response options).

Category and data element		FHIR resource	Items (N=48), n
Anamnesis			
	Chemotherapy	Procedure	1
	Immunosuppressive therapy	Procedure	1
	Regular alcohol intake	Observation	2
COVID-19 infection and treatment			
	Disease course	Encounter, Procedure	5
	SARS-CoV-2 infection	Condition	1
	SARS-CoV-2 variant	Observation	1
Immunization			
	Contraindications to immunization	Immunization	2
	Immunizations performed	Immunization	3
	Reason for immunization	Immunization	5
	Willingness to receive additional immunization doses	Observation	1
Immunization reactions			
	Analgesic or antipyretic drug intake	MedicationStatement	1
	Body temperature	Observation	1
	Complications after immunization	Observation	5
	Medical treatment for adverse reactions	Encounter	3
	Symptoms after vaccination	Condition	16

[a]GECCO: German Corona Consensus.

[b]FHIR: Fast Healthcare Interoperability Resources.

Table . Types of data elements in the pediatrics extension module extending the GECCO[a] core data set. Shown are the data elements and the FHIR[b] resource they have been mapped to, as well as the number of items for each data element (ie, different response options).

Category and data element		FHIR resource	Items (N=150), n
Complications			
	Complications to COVID-19	Condition	47
Demographics			
	Body measures	Observation	6
Imaging			
	Echocardiography	Procedure, ImagingStudy	1
	PET-CT[c]	Procedure, ImagingStudy	1
Immunization			
	Immunizations performed	Immunization	2
Laboratory values			
	Laboratory values	Observation	27
Medical history			
	Chronic hematologic diseases	Condition	8
	Chronic kidney diseases	Condition	2
	Congenital disease	Condition	1
	Gastrointestinal diseases	Condition	6
	Medical history stem cells transplant	Condition	2
Medication			
	Medication	MedicationStatement, List	35
Symptoms			
	COVID-19 symptoms	Condition	9
Therapy			
	Hospitalization	Observation	2
	Thoracic drainage	Procedure	1

[a]GECCO: German Corona Consensus.

[b]FHIR: Fast Healthcare Interoperability Resources.

[c]PET-CT: positron emission tomography–computed tomography.

Table . Types of data elements in the cardiology extension module extending the GECCO[a] core data set. Shown are the data elements and the FHIR[b] resource they have been mapped to, as well as the number of items for each data element (ie, different response options).

Category and data element		FHIR resource	Items (N=52), n
Anamnesis			
	Chronic cardiologic diseases	Condition	6
COVID-19−related complications			
	Cardiologic complications of COVID-19	Condition	7
Echocardiography			
	Echocardiography findings	Observation	20
	Echocardiography procedure	Procedure	3
Electrocardiography			
	Electrocardiography findings	Observation	11
	Electrocardiography procedure	Procedure	2
Laboratory values			
	Laboratory values	Observation	2
Medication			
	Angiotensin receptor antagonist	MedicationStatement	1

[a]GECCO: German Corona Consensus.

[b]FHIR: Fast Healthcare Interoperability Resources.

All data items were mapped to the appropriate FHIR resources (Observation, Condition, Procedure, MedicationStatement, Encounter, Questionnaire, QuestionnaireResponse, Immunization, ImagingStudy, List, and Specimen), and 26, 14, and 18 profiles (25, 17, and 12 value sets) were created for the immunization, pediatrics, and cardiology extension modules, respectively. The data items that were already part of the GECCO data set and not removed during the data selection step were taken from the GECCO data set and referenced as such in the implementation guides.

The implementation guides for the three extension modules have been published on GitHub pages [28-30]. The source FSH files have been published on GitHub [31-33]. Logical models and data set descriptions are hosted on ART-DECOR, an open collaboration platform for modeling data set definitions, their descriptions, and their terminology bindings [34-36].

Discussion

Principal Findings

We herein present an interdisciplinary, iterative, consensus-based workflow for the definition of research data sets, focusing on creating data sets with the most relevant data elements for a particular field of study and on creating universally usable data sets according to the FAIR principles [22]. We applied the workflow to develop 3 GECCO extension modules that contain data items that are relevant for COVID-19–related patient research on infectious diseases (with a focus on immunization), pediatrics, and cardiology. These extension modules complement the GECCO core data set for domain-specified research. The data items are represented in HL7 FHIR profiles and use international terminologies to ensure a harmonized, standardized, and interoperable data set definition for these medical domains. The provision of data according to the extension modules introduced in this paper will enable cross-institutional and cross-country data collection and collaborative research with a particular focus on immunization, pediatrics, and cardiology.

We have specified and implemented an interdisciplinary, iterative, consensus-based workflow for the selection of data items and the development of the data set definition. Close collaboration and constant feedback loops with domain experts from various medical specialties right from the beginning of a project, as performed in this study, are key for the successful development of a useful data set definition. Indeed, since the selection of relevant data items in this study was driven by the end users of the data set, who are the researchers that later will be using the data for their specialized areas of research, the semantic usability of the data sets is guaranteed. Likewise, having medical information specialists develop the formal data set specification ensures the technical interoperability and usability of the data set definition. In this study, we focused on the initial development of interoperable data set definitions for COVID-19–related patient research on infectious diseases (with a focus on immunization), pediatrics, and cardiology. To ensure the continued accuracy and relevance of the data set definitions, such data set definitions should be regularly subjected to public review and balloting processes following the initial development. For example, a revised version of the GECCO data set will undergo HL7 balloting, pending stakeholders' approval.

Although general interoperability in health care and clinical research is difficult to achieve, we focused on achieving syntactic and semantic interoperability of the data set definitions,

which are 2 of the 4 levels into which interoperability can be distinguished, alongside technical and organizational interoperability [8]. We pursued semantic interoperability by using international standardized vocabularies, such as those provided by the LOINC and SNOMED CT vocabularies, to ensure that the meanings of the data elements and their interpretations were unambiguous. We pursued syntactic interoperability by using an open standard for data representation, namely the HL7 FHIR standard, which provides a flexible and extensible framework for exchanging data elements and resources between different systems and applications. We did not focus on organizational interoperability in our work, as this requires coordination and alignment between different health care organizations and stakeholders, which can be challenging in practice. Although we did not specifically address organizational interoperability in our study, we believe that our approach to achieving semantic and syntactic interoperability can contribute to broader efforts toward achieving organizational interoperability over time.

In addition to the successful development of data set definitions, several factors determine a successful deployment or the use of the developed extension modules [37]. First and most importantly, clear and concise documentation of how to implement and provide data using the data set definition is required. For FHIR-based data set definitions, so-called *implementation guides* are used to provide a narrative overview as well as technical details on the data set definition [38]. Thus, we have created and published implementation guides for each of the here developed extension modules. Second, the example implementations of the extension modules serve as a blueprint for developers and data engineers who implement the extension modules for their clinical databases. From our experience with the implementation of the GECCO data set, well-defined example data items may be of equal if not higher importance than the technical description of the data set specification, as developers and engineers tend to use the examples as blueprints for their implementation. Thus, we equipped every FHIR profile defined in the extension modules with at least one example. These examples are incorporated and issued within the implementation guides of the modules. Specifically, we aimed to provide 1 example for each different category of response option per profile. Third, the actual implementation of the extension modules should be part of follow-up infrastructure projects to supply funding and resources for filling the data set definition with actual data. For the GECCO data set, this is ensured by follow-up projects of the German COVID-19 Research Network of University Medicine (“Netzwerk Universitätsmedizin”), such as CODEX+ (Collaborative Data Exchange and Usage), which includes several implementation tasks that are actively using the GECCO data set items [39] and further projects [40-43]. Fourth, once the data set definitions are implemented and leveraged in use cases, additional demands to the data set are likely raised, or issues with existing definitions are revealed. The maintenance of existing definitions (eg, performing technical corrections, evolving the definitions, or adding new items) is, therefore, necessary and must be organized and funded. Last, successful use of the extension modules is also highly dependent on the degree of interoperability of the data set definitions [1,44,45]. For example, the use of questionnaires to assess certain features is common in clinical research. However, depending on the exact wording of the question and the number and wording of response options, results from different studies might not be directly comparable even if they assessed the same features, as the questions and response options differ between studies. In the presented extension modules, several items were at first specified in a questionnaire-like fashion, and the direct implementation of these as Questionnaire resources in FHIR would have limited the applicability of such data elements, especially when aiming to map these elements from an electronic health record system. In these cases, we revised the data element specification to use interoperable concepts rather than questions. Here, repeated consultation with and final approval of the group of medical experts were key to being able to convert questions into interoperable concepts that convey the same information as that intended by the content definition of the group of medical experts. In general, we recommend not to use Questionnaire and QuestionnaireResponse FHIR profiles in cases where the information to be represented can be modeled by using more general, interoperable concepts and FHIR resources.

The challenges of creating and harmonizing COVID-19 data sets are not unique to our work, and although initiatives, such as the Clinical Data Interchange Standards Consortium (CDISC), have released guidance on how to represent COVID-19 research data in a standardized format [46], the actual selection of the relevant biomedical concepts to be represented is left to the implementers. We explicitly selected the data elements for COVID-19–related patient research that are the most relevant for further characterizing patients with respect to research in infectious diseases (with a focus on immunization), pediatrics, and cardiology. However, we recognize the need for ongoing collaboration and standardization efforts to maximize interoperability and facilitate data sharing and analysis. Such efforts include integrating the GECCO data set with other COVID-19–related data sets and standards, both within and between countries. For example, we are currently harmonizing the GECCO data set with the ORCHESTRA (Connecting European Cohorts to Increase Common and Effective Response to SARS-CoV-2 Pandemic) project, which intends to create a harmonized and standardized data set for a pan-European cohort for COVID-19 research [40]. To facilitate the mapping of the data items that were developed in our work and represented in HL7 FHIR to the CDISC Study Data Tabulation Model standard, the organizations behind the two standards have collaboratively developed a comprehensive implementation guide, thereby enabling mapping between the different standards, ensuring compatibility, and facilitating interoperability across systems [47]. Moving forward, we encourage developers of tools and resources to facilitate the mapping and harmonization of different data standards, and we look forward to continued collaboration with the wider research community to address these challenges and advance COVID-19 research.

Conclusion

We herein introduce the development workflow and the resulting data set definitions for GECCO extension modules for the immunization, pediatrics, and cardiology domains. We have defined and implemented a workflow in which interdisciplinary

teams of medical domain experts, medical information specialists, and FHIR developers closely collaborate in an iterative, consensus-based fashion for the successful development of useful and interoperable data set definitions. This workflow may serve as a blueprint for further data set definition projects, such as the further development of data set definitions for extending the GECCO core data set. The extension modules described in this work have been validated and published. Their implementation and active use are anticipated in the context of current nationwide COVID-19 research networks in Germany.

Acknowledgments

The NAPKON ("Nationales Pandemie Kohorten Netz"; German National Pandemic Cohort Network) project is funded under a scheme issued by the Network of University Medicine (Nationales Forschungsnetzwerk der Universitätsmedizin [NUM]) by the Federal Ministry of Education and Research of Germany (Bundesministerium für Bildung und Forschung [BMBF]; grant number 01KX2021). The funding body did not take a role in the design of the study, in the development of the data set, or in the writing of the manuscript. We thank Yannick Börner for his valuable contribution to the definition of the Fast Healthcare Interoperability Resources (FHIR) profiles. We thank all members of the subject- and organ-specific working groups.

Data Availability

The implementation guides for the three extension modules have been published on GitHub pages [28-30]. The source Fast Healthcare Interoperability Resources Shorthand (FSH) files have been published on GitHub [31-33]. Data set descriptions can be found on ART-DECOR [34-36].

Authors' Contributions

All authors contributed to the development of the extension modules. GL, TH, SB, LR, JS, AB, and ST performed terminology mapping, FHIR profiling, and critical review of the concept and resource mappings. TH, SB, and LR defined the data sets in ART-DECOR. DH, FK, LES, FE, NT, RB, AF, and MD developed and compiled the list of data items for the data sets. SR, LL, and MU coordinated the project and the consensus finding process within and between working groups. JJV, CvK, and ST conceived the work. GL drafted the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

ST is the vice chair of Health Level Seven International (HL7) Germany. The other authors declare that they have no competing interests.

Multimedia Appendix 1
Supplementary tables.
[PDF File, 283 KB - medinform_v11i1e45496_app1.pdf]

References

1. Lehne M, Sass J, Essenwanger A, Schepers J, Thun S. Why digital medicine depends on interoperability. NPJ Digit Med 2019 Aug 20;2:79. [doi: 10.1038/s41746-019-0158-1] [Medline: 31453374]
2. Sass J, Bartschke A, Lehne M, Essenwanger A, Rinaldi E, Rudolph S, et al. The German Corona Consensus Dataset (GECCO): a standardized dataset for COVID-19 research in university medicine and beyond. BMC Med Inform Decis Mak 2020 Dec 21;20(1):341. [doi: 10.1186/s12911-020-01374-w] [Medline: 33349259]
3. Gruendner J, Deppenwiese N, Folz M, Köhler T, Kroll B, Prokosch HU, et al. The architecture of a feasibility query portal for distributed COVID-19 Fast Healthcare Interoperability Resources (FHIR) patient data repositories: design and implementation study. JMIR Med Inform 2022 May 25;10(5):e36709. [doi: 10.2196/36709] [Medline: 35486893]
4. Sedlmayr B, Sedlmayr M, Kroll B, Prokosch HU, Gruendner J, Schüttler C. Improving COVID-19 research of university hospitals in Germany: formative usability evaluation of the CODEX feasibility portal. Appl Clin Inform 2022 Mar;13(2):400-409. [doi: 10.1055/s-0042-1744549] [Medline: 35445386]
5. Millar J. The need for a global language – SNOMED CT introduction. Stud Health Technol Inform 2016;225:683-685. [doi: 10.3233/978-1-61499-658-3-683] [Medline: 27332304]
6. Fiebeck J, Gietzelt M, Ballout S, Christmann M, Fradziak M, Laser H, et al. Implementing LOINC - current status and ongoing work at a medical university. Stud Health Technol Inform 2019 Sep 3;267:59-65. [doi: 10.3233/SHTI190806] [Medline: 31483255]
7. McDonald CJ, Huff SM, Suico JG, Hill G, Leavelle D, Aller R, et al. LOINC, a universal standard for identifying laboratory observations: a 5-year update. Clin Chem 2003 Apr;49(4):624-633. [doi: 10.1373/49.4.624] [Medline: 12651816]
8. Lehne M, Luijten S, Vom Felde Genannt Imbusch P, Thun S. The use of FHIR in digital health - A review of the scientific literature. Stud Health Technol Inform 2019 Sep 3;267:52-58. [doi: 10.3233/SHTI190805] [Medline: 31483254]

9. Vorisek CN, Lehne M, Klopfenstein SAI, Mayer PJ, Bartschke A, Haese T, et al. Fast Healthcare Interoperability Resources (FHIR) for interoperability in health research: systematic review. JMIR Med Inform 2022 Jul 19;10(7):e35724. [doi: 10.2196/35724] [Medline: 35852842]
10. Jacobsen A, Kaliyaperumal R, da Silva Santos LOB, Mons B, Schultes E, Roos M, et al. A generic workflow for the data FAIRification process. Data Intell 2020 Jan 1;2(1-2):56-65. [doi: 10.1162/dint_a_00028]
11. SNOMED International. URL: www.snomed.org/ [accessed 2022-03-16]
12. Bundesinstitut Für Arzneimittel und Medizinprodukte (Bfarm). ICD-10-GM. URL: www.bfarm.de/EN/Code-systems/Classifications/ICD/ICD-10-GM/_node.html [accessed 2022-03-16]
13. Bundesinstitut Für Arzneimittel und Medizinprodukte (Bfarm). ATC. URL: www.bfarm.de/DE/Kodiersysteme/Klassifikationen/ATC/_node.html [accessed 2022-03-16]
14. Regenstrief Institute. UCUM. URL: ucum.org/trac [accessed 2022-03-16]
15. art-decor.org. URL: www.art-decor.org/mediawiki/index.php/Main_Page [accessed 2022-03-16]
16. SIMPLIFIER.NET. Medizininformatik Initiative. URL: simplifier.net/organization/koordinationsstellemii [accessed 2022-03-15]
17. SIMPLIFIER.NET. Kassenärztliche Bundesvereinigung (KBV). URL: simplifier.net/organization/kassenrztlichebundesvereinigungkbv [accessed 2022-03-15]
18. HL7 International. FHIR shorthand. URL: hl7.org/fhir/uv/shorthand/ [accessed 2022-04-25]
19. GitHub. SUSHI unshortens short hand inputs. 2022. URL: github.com/FHIR/sushi [accessed 2022-04-25]
20. Lichtner G. GitHub. FHIR shorthand validator. 2021. URL: github.com/glichtner/fsh-validator [accessed 2022-03-15]
21. GitHub. napkon-module-template. 2022. URL: github.com/BIH-CEI/napkon-module-template [accessed 2022-03-15]
22. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016 Mar 15;3:160018. [doi: 10.1038/sdata.2016.18] [Medline: 26978244]
23. HL7 International. Implementation guide registry. URL: fhir.org/guides/registry/ [accessed 2022-11-3]
24. Schons M, Pilgram L, Reese JP, Stecher M, Anton G, Appel KS, et al. The German National Pandemic Cohort Network (NAPKON): rationale, study design and baseline characteristics. Eur J Epidemiol 2022 Aug;37(8):849-870. [doi: 10.1007/s10654-022-00896-z] [Medline: 35904671]
25. Hillus D, Schwarz T, Tober-Lau P, Vanshylla K, Hastor H, Thibeault C, et al. Safety, Reactogenicity, and Immunogenicity of Homologous and heterologous prime-boost Immunisation with Chadox1 nCoV-19 and BNT162b2: a prospective cohort study. Lancet Respir Med 2021;9(11):1255-1265. [doi: 10.1016/S2213-2600(21)00357-X] [Medline: 34391547]
26. Tober-Lau P, Schwarz T, Vanshylla K, Hillus D, Gruell H, EICOV/COVIM Study Group, et al. Long-term immunogenicity of BNT162b2 vaccination in older people and younger health-care workers. Lancet Respir Med 2021 Sep;9(11):e104-e105. [doi: 10.1016/S2213-2600(21)00456-2] [Medline: 34687656]
27. COVIM. COVIM – COllaboratiVe IMmunity Platform of the NUM. URL: covim-netzwerk.de/ [accessed 2022-03-16]
28. NAPKON. NAPKON cardiology module. URL: bih-cei.github.io/napkon-cardiology/ [accessed 2022-03-15]
29. NAPKON. NAPKON pediatrics module. URL: bih-cei.github.io/napkon-pediatrics/ [accessed 2022-03-15]
30. NAPKON. NAPKON vaccination module. URL: bih-cei.github.io/napkon-vaccination/ [accessed 2022-03-15]
31. GitHub. NAPKON cardiology module. 2022. URL: github.com/BIH-CEI/napkon-cardiology [accessed 2022-03-15]
32. GitHub. NAPKON pediatrics module. 2022. URL: github.com/BIH-CEI/napkon-pediatrics [accessed 2022-03-15]
33. Github. NAPKON vaccination module. 2022. URL: github.com/BIH-CEI/napkon-vaccination [accessed 2022-03-15]
34. art-decor.org. NAPKON cardiology module. URL: art-decor.org/art-decor/decor-datasets--covid19f-?id=2.16.840.1.113883.3.1937.777.53.1.2&effectiveDate=2020-08-12T00%3A00%3A00&conceptId=2.16.840.1.113883.3.1937.777.53.2.250&conceptEffectiveDate=2021-02-16T13%3A25%3A43&language=en-US [accessed 2022-03-16]
35. art-decor.org. NAPKON pediatrics module. URL: art-decor.org/art-decor/decor-datasets--covid19f-?id=2.16.840.1.113883.3.1937.777.53.1.2&effectiveDate=2020-08-12T00%3A00%3A00&conceptId=2.16.840.1.113883.3.1937.777.53.2.12&conceptEffectiveDate=2020-09-18T09%3A20%3A12&language=en-US [accessed 2022-03-16]
36. art-decor.org. NAPKON vaccination module. URL: art-decor.org/art-decor/decor-datasets--covid19f-?id=2.16.840.1.113883.3.1937.777.53.1.2&effectiveDate=2020-08-12T00%3A00%3A00&conceptId=2.16.840.1.113883.3.1937.777.53.2.453&conceptEffectiveDate=2021-08-25T12%3A45%3A26&language=en-US [accessed 2022-03-16]
37. Kush RD, Warzel D, Kush MA, Sherman A, Navarro EA, Fitzmartin R, et al. FAIR data sharing: the roles of common data elements and harmonization. J Biomed Inform 2020 Jul;107:103421. [doi: 10.1016/j.jbi.2020.103421] [Medline: 32407878]
38. Shivers J, Amlung J, Ratanaprayul N, Rhodes B, Biondich P. Enhancing narrative clinical guidance with computer-readable artifacts: Authoring FHIR implementation guides based on WHO recommendations. J Biomed Inform 2021 Oct;122:103891. [doi: 10.1016/j.jbi.2021.103891] [Medline: 34450285]
39. Lichtner G, Alper BS, Jurth C, Spies C, Boeker M, Meerpohl JJ, et al. Representation of evidence-based clinical practice guideline recommendations on FHIR. J Biomed Inform 2023 Mar;139:104305. [doi: 10.1016/j.jbi.2023.104305] [Medline: 36738871]

40. Rinaldi E, Stellmach C, Rajkumar NMR, Caroccia N, Dellacasa C, Giannella M, et al. Harmonization and standardization of data for a pan-European cohort on SARS-CoV-2 pandemic. NPJ Digit Med 2022 Jun 14;5(1):75. [doi: 10.1038/s41746-022-00620-x] [Medline: 35701537]
41. Mang JM, Prokosch HU, Kapsner LA. Reproducibility in 2023 - an end-to-end template for analysis and manuscript writing. Stud Health Technol Inform 2023 Jun 18;302:58-62. [doi: 10.3233/SHTI230064] [Medline: 37203609]
42. Horn A, Krist L, Lieb W, Montellano FA, Kohls M, Haas K, et al. Long-term health sequelae and quality of life at least 6 months after infection with SARS-CoV-2: design and rationale of the COVIDOM-study as part of the NAPKON population-based cohort platform (POP). Infection 2021 Dec;49(6):1277-1287. [doi: 10.1007/s15010-021-01707-5] [Medline: 34642875]
43. Prokosch HU, Bahls T, Bialke M, Eils J, Fegeler C, Gruendner J, et al. The COVID-19 data exchange platform of the German University Medicine. Stud Health Technol Inform 2022 May 25;294:674-678. [doi: 10.3233/SHTI220554] [Medline: 35612174]
44. Perlin JB. Health information technology Interoperability and use for better care and evidence. JAMA 2016 Oct 25;316(16):1667-1668. [doi: 10.1001/jama.2016.12337] [Medline: 27669026]
45. Cutillo CM, Sharma KR, Foschini L, Kundu S, Mackintosh M, Mandl KD, et al. Machine intelligence in healthcare—perspectives on trustworthiness, explainability, usability, and transparency. NPJ Digit Med 2020 Mar 26;3(1):47. [doi: 10.1038/s41746-020-0254-2] [Medline: 32258429]
46. Clinical Data Interchange Standards Consortium. COVID-19 therapeutic area user guide v1.0. 2021. URL: www.cdisc.org/standards/therapeutic-areas/covid-19/covid-19-therapeutic-area-user-guide-v1-0 [accessed 2023-03-10]
47. Clinical Data Interchange Standards Consortium. FHIR to CDISC joint mapping implementation guide v1.0. 2021. URL: www.cdisc.org/standards/real-world-data/fhir-cdisc-joint-mapping-implementation-guide-v1-0 [accessed 2023-03-10]

Abbreviations

CDISC: Clinical Data Interchange Standards Consortium
CODEX+: Collaborative Data Exchange and Usage
COVIM: Collaborative Immunity Platform of the Netzwerk Universitätsmedizin
FAIR: Findable, Accessible, Interoperable, Reusable
FHIR: Fast Healthcare Interoperability Resources
FSH: Fast Healthcare Interoperability Resources Shorthand
GECCO: German Corona Consensus
HL7: Health Level Seven International
LOINC: Logical Observation Identifiers Names and Codes
NAPKON: Nationales Pandemie Kohorten Netz
ORCHESTRA: Connecting European Cohorts to Increase Common and Effective Response to SARS-CoV-2 Pandemic
SNOMED CT: Systematized Nomenclature of Medicine-Clinical Terms

Edited by J Klann; submitted 04.01.23; peer-reviewed by F Amar, S Hume, S Sarbadhikari; revised version received 16.03.23; accepted 04.04.23; published 18.07.23.

Please cite as:
Lichtner G, Haese T, Brose S, Röhrig L, Lysyakova L, Rudolph S, Uebe M, Sass J, Bartschke A, Hillus D, Kurth F, Sander LE, Eckart F, Toepfner N, Berner R, Frey A, Dörr M, Vehreschild JJ, von Kalle C, Thun S
Interoperable, Domain-Specific Extensions for the German Corona Consensus (GECCO) COVID-19 Research Data Set Using an Interdisciplinary, Consensus-Based Workflow: Data Set Development Study
JMIR Med Inform 2023;11:e45496
URL: https://medinform.jmir.org/2023/1/e45496
doi:10.2196/45496

Original Paper

Data Analysis of Physician Competence Research Trend: Social Network Analysis and Topic Modeling Approach

So Jung Yune[1*], PhD; Youngjon Kim[2*], PhD; Jea Woog Lee[3*], PhD

[1]Department of Medical Education, Pusan National University, Busan, Republic of Korea

[2]Department of Medical Education, Wonkwang University School of Medicine, Iksan, Republic of Korea

[3]Intelligence Informatics Processing Lab, Chung-Ang University, Seoul, Republic of Korea

[*]all authors contributed equally

Corresponding Author:
Jea Woog Lee, PhD
Intelligence Informatics Processing Lab
Chung-Ang University
84, Heukseok-ro, Dongjak-gu,
Seoul, 06974
Republic of Korea
Phone: 82 10 5426 7318
Email: yyizeuks@cau.ac.kr

Related Article:

This is a corrected version. See correction statement: https://medinform.jmir.org/2023/1/e53484

Abstract

Background: Studies on competency in medical education often explore the acquisition, performance, and evaluation of particular skills, knowledge, or behaviors that constitute physician competency. As physician competency reflects social demands according to changes in the medical environment, analyzing the research trends of physician competency by period is necessary to derive major research topics for future studies. Therefore, a more macroscopic method is required to analyze the core competencies of physicians in this era.

Objective: This study aimed to analyze research trends related to physicians' competency in reflecting social needs according to changes in the medical environment.

Methods: We used topic modeling to identify potential research topics by analyzing data from studies related to physician competency published between 2011 and 2020. We preprocessed 1354 articles and extracted 272 keywords.

Results: The terms that appeared most frequently in the research related to physician competency since 2010 were *knowledge*, *hospital*, *family*, *job*, *guidelines*, *management*, and *communication*. The terms that appeared in most studies were *education*, *model*, *knowledge*, and *hospital*. Topic modeling revealed that the main topics about physician competency included *Evidence-based clinical practice*, *Community-based healthcare*, *Patient care*, *Career and self-management*, *Continuous professional development*, and *Communication and cooperation*. We divided the studies into 4 periods (2011-2013, 2014-2016, 2017-2019, and 2020-2021) and performed a linear regression analysis. The results showed a change in topics by period. The *hot topics* that have shown increased interest among scholars over time include *Community-based healthcare*, *Career and self-management*, and *Continuous professional development*.

Conclusions: On the basis of the analysis of research trends, it is predicted that physician professionalism and community-based medicine will continue to be studied in future studies on physician competency.

(JMIR Med Inform 2023;11:e47934) doi:10.2196/47934

KEYWORDS

physician competency; research trend; competency-based education; professionalism; topic modeling; latent Dirichlet allocation; LDA algorithm; data science; social network analysis

Introduction

Background

Medical publications began defining competencies in the 1970s [1-3]. Physician competency refers to the essential qualities that a physician should possess. The search for physician competency begins with the question of what it means to be a physician. Competency connotes various ideas, such as which physician traits are desired by society and what supports and promotes this transition of identity. Competency entails the concept of the physician as a professional, what the physician can do, and how the physician approaches their practice [4]. In summary, competency is considered a complex set of behaviors built on the components of knowledge, skills, attitudes, and competence as a personal ability [5]. Competency in this study is the core competency required to successfully perform a physician's job and includes knowledge, skills, and attitude, regardless of the specific major.

In the medical profession, the competency theme began by pursuing ways for the medical circle to improve the performance of health care workers. The issue arose in response to the growing demands of medical consumers and society as a neoliberalistic ideology spread in the 1970s and consumers' awareness of their rights increased. In 1972, the American Academy of Pediatrics discussed physician competency by publishing a foundation for evaluating pediatricians' competency [3]. In 1978, the World Health Organization found the cause of declines in medical service quality to be inadequate education and attempted to improve health care providers' competencies through Competency-Based Curriculum Development in Medical Education [6]. Since then, this approach has influenced medical education globally, starting with the Royal College of Physicians and Surgeons in Canada and the Accreditation Council for Graduate Medical Education in the United States. Similar programs have been established worldwide, influencing strategies for global human resources and international partnerships for medical training [7]. These include the Outcome Project of the US Accreditation Council for Graduate Medical Education, General Medical Council's Tomorrow's Doctors [8-10], Scottish Doctor [11], and Canadian CanMEDS framework [12].

For physicians, competency varies depending on the clinical, cultural, and geographical context [13]. In medical practice, the perception of a medical professional's competent role changes continuously over time. In the early 19th century, physicians applied ointments and drew blood but did not deliver babies. In the 21st century, physicians are required to use advanced technologies and artificial intelligence in medical surgeries [14,15]. The ability to use technologies required by future health care systems is a challenge for physicians. However, at the same time, communication with patients and colleagues and interprofessional teamwork are essential human skills, and personal traits, such as empathy, humility, compassion, emotional intelligence, and a passion for continuous learning are also emphasized. Varying levels of health care infrastructure over time [16]; social awareness of minority groups [17]; and occasional health care challenges, such as global pandemics [18], emphasize certain specific physician competencies. Physicians' competency, initially discussed within the scope of their social accountability, includes changes in their roles from the perspective of patients, health care, and self-management [19] of their health and wellness [20]. Thus, physicians' abilities reflect their social situation and demands. In addition, physicians' core abilities are expected to change over time. Therefore, this study, which analyzed changes in physician competency by period, will help understand the social demands expected of physicians in each period and identify research topics that should be important in medical education in the future.

To date, studies on physician competency have focused on literature review [21-23] or meta-analyses [24,25]. However, systematic literature reviews have limitations in deriving comprehensively synthesized results because their analyses focus on narrow areas such as subjects, variables, environment, and intervention. In medical education, studies on competency often explore the acquisition, performance, and evaluation of particular skills, knowledge, or behaviors that constitute physician competency. Some of these studies examined patient-physician communication [26], risk management for emergency physicians, technical skills in robotic surgery in urological practice [27], and the instruction of medical staff [28]. Nevertheless, such approaches fail to convey the trends in physician competency research because they explore the essential medical skills for a specific task in a certain context. In addition, people may have different thoughts about the core competencies that physicians should possess. For example, a patient's expectations of a physician's ability may differ from a senior physician's expectation of a junior physician's ability. Therefore, a more macroscopic method is required to analyze the core competencies of physicians in this era.

Recently, new big data analysis techniques such as social network analysis and topic modeling have been used. These approaches have the advantage of organizing the knowledge structure context by forecasting the trajectory of change in research and future issues and revealing the correlation between concepts. Social network analysis involves detecting influential core keywords in a vast amount of text and showing the relationship between keywords, allowing researchers to comprehend the context intuitively [29,30]. Topic modeling detects hidden topics in text data, analyzes the association and distribution of each topic, and provides integrated information [31]. From a microscopic perspective, it identifies core topics and their relationships. From a macroscopic perspective, it identifies the flow and context of core topics and the trend of topics by period [32]. Text network analysis and topic modeling are ideal approaches for analyzing trends in research on physicians' competencies and accomplishing research objectives.

Objectives

The research problems based on the abovementioned research necessity are as follows:

1. Extract core keywords from physician competency studies and create a network

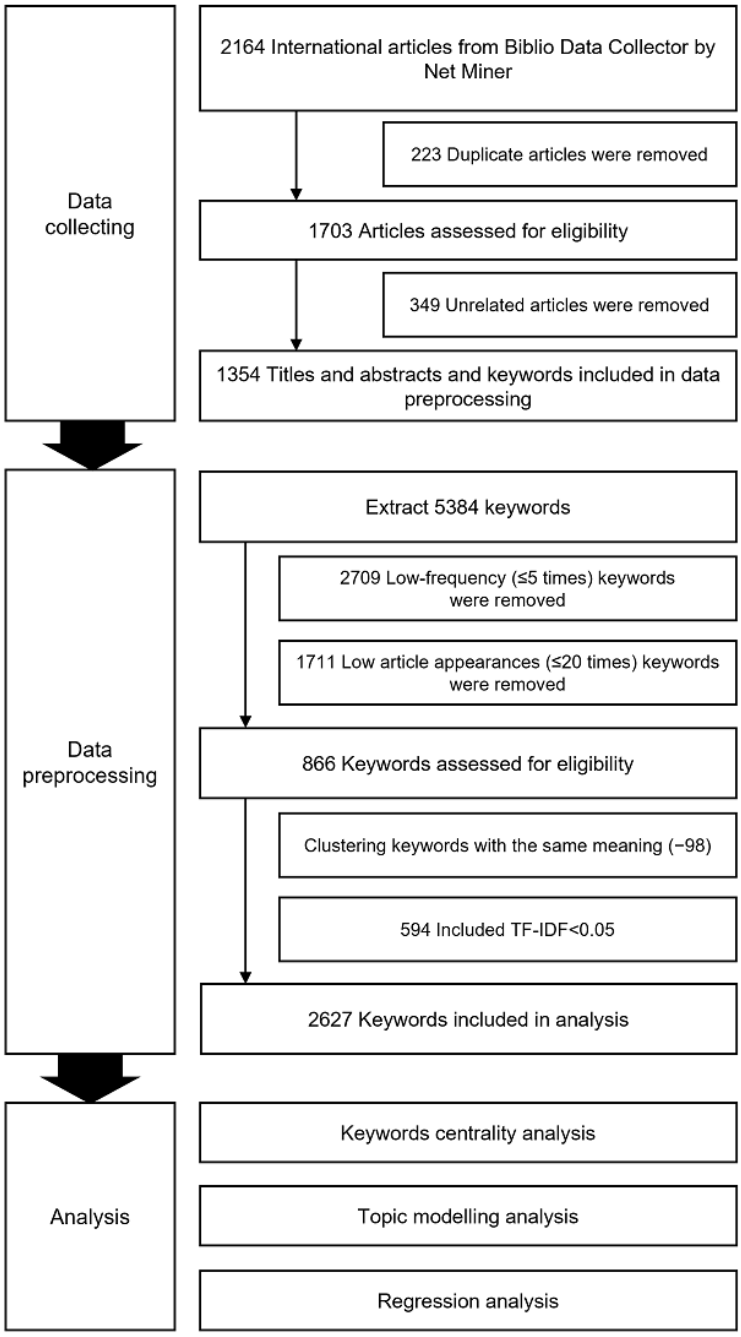
2. Examine the structure and characteristics of the network created based on physician competency studies
3. Examine the main topics through topic modeling in physician competency studies
4. Examine the trend of physician competency studies by topic based on time flow

Methods

Data Collection

To examine the flow of academic research on physician competency, we selected related research articles as raw data, which were collected using NetMiner Biblio Data Collector (Cyram Inc). We selected 10 years (2011-2020) and collected 2164 articles on physician competency published in *Springer* using the following keywords: "(doctors or physicians) AND ((competence or competency or competencies) or (expertise or expert or proficiency) or (responsibility or accountability or liability or blameworthiness) or (profession or occupation or roles or duties or jobs) or (performance and practice) or professionalism))." After eliminating 810 articles that overlapped or were unrelated, 1354 articles remained for our analysis. The research flow is illustrated in Figure 1.

Figure 1. Flow of research procedures. TF-IDF: term frequency–inverse document frequency.



Data Preprocess

Because we could not use the original text in the preselected articles in the analysis, we processed the sentences into separate words as units of analysis. The collected papers consisted of natural language sentences such as theories, knowledge, and opinions. However, the sentences cannot be used directly in the analysis. Therefore, steps must be taken to convert each sentence into an individual word that can be analyzed [33]. In this study, to extract nouns, adjectives, and verbs, we used the morpheme-refining function of the NetMiner program and extracted 5384 words from the titles, abstracts, and keywords of the research articles. Preprocessing was performed to convert these words into analyzable keyword data.

Nouns that were unsuitable for analysis were eliminated. First, we removed words that appeared 5 times or keywords that did not appear in more than 20 of the 1354 articles. In the final stage, we eliminated infrequently appearing words (<5 times) and extremely common words that appeared frequently in all papers (term frequency–inverse document frequency<0.05) [34]. After preprocessing, 272 words were extracted.

Social Network Analysis

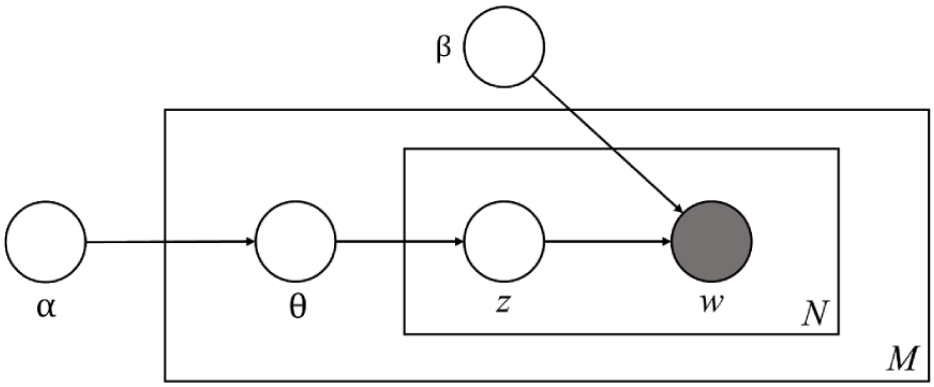
In this study, social network analysis was used to examine the knowledge structure and characteristics through keyword extraction and network generation from physician competency studies, designated as research problems 1 and 2. In addition, the roles of keywords in the network were determined by assessing their importance using social network analysis techniques such as degree centrality, closeness centrality, and betweenness centrality. Social network analysis determines core nodes based on degree centrality, closeness centrality, and betweenness centrality. The degree centrality of a node increases with the number of nodes directly connected to it. Thus, the degree centrality indicates the influence of a node (keyword) based on the number of connected nodes. The betweenness centrality indicates the centrality of a node (keyword) between 2 other nodes (keywords). Betweenness centrality increases when the number of times a node appears on the shortest path between 2 other nodes increases. Keywords with high betweenness centrality control the information flow and exert a substantial influence on the overall connectivity of the network. To visually understand the positions and relationships among keywords, we used spring mapping from NetMiner 4.0. Spring mapping maximizes the characteristics of branching out the graph by placing the connected nodes closer and the disconnected nodes farther, using a simulated annealing technique to balance the 2 forms of distance.

Topic Modeling

To address research problem 3, we conducted topic modeling to explore the topic areas of physician competency research. Topic modeling was performed using keywords extracted from social network analysis. NetMiner 4.0 was used as the analysis program. Topic modeling is used to predict latent topics based on the association of multiple words in a text. Topic modeling extracts topics from research papers through keyword exploration, which helps recognize knowledge structures and patterns [35]. Knowledge structures are defined as a visualization of keyword clustering and a network map of how concepts in domain knowledge are interrelated [36]. This pattern is defined as the process of change in the knowledge structure over a period [37]. Topic modeling, as a big data analysis technique, provides a quantitative approach for identifying previously undiscovered macrotopic areas in physician competency research. Topics were extracted by applying a latent Dirichlet allocation (LDA) algorithm. The LDA algorithm is a probability model for predicting hidden topics by analyzing the distribution of words observed in a document. It is useful for reducing the data size and producing consistent topics.

Figure 2 shows a graphical model of LDA. The boxes in the figure are “plates,” which represent duplicates. The outer plate represents the document (M) and the inner plate represents repeatedly selected topics (z) and words (w) within the document (N). “ ” indicates the distribution of topics in the document. Both “α” and “β” are hyperparameters that indicate Dirichlet distribution. LDA cannot be directly used to determine the number of topics, so the third hyperparameter is the “number of topics” the algorithm will discover. The probability calculation formula is as follows [38]:

Figure 2. Illustration for the conceptual model of latent Dirichlet allocation. α: a parameter that represents the Dirichlet prior for the document topic distribution; β: a parameter that represents the Dirichlet for the word distribution; θ: a vector for topic distribution over a document d; z: a topic for a chosen word in a document; w: specific words in N; M: document length; N: number of words in the document.



To increase the accuracy of the results, appropriate Cronbach α values, β values, number of topics, and keywords should be determined [39]. To evaluate how well each keyword described each topic, we used one of the topic consistency metrics, the silhouette coefficient. The silhouette coefficient is an indicator that evaluates how well keywords, which are components of a topic, are classified [40]. A value closer to 1 indicates that the keywords within each cluster are well formed. Moreover, good clustering means that similarity to other topics is low and keywords within the same cluster describe a topic well. For topic t characterized by a higher-order word set W_t (any word whose probability exceeds a predefined threshold or a fixed number of high-order words), the consistency formula is defined as follows [41]:

In this study, we set the term frequency–inverse document frequency threshold value at 0.5 and word length at 2. We used a silhouette coefficient to calculate the optimal values for Cronbach α, β, and the number of topics. A silhouette coefficient (or score) closer to 1 had higher explanatory power, validating Cronbach α, validating β, and the number of topics and descriptions of the object in the topic model were well matched. We also used a silhouette-clustering configuration. To determine the optimal number of topics, we conducted a comprehensive analysis by varying the number of topics from 5 to 30 and exploring Cronbach α values ranging from .01 to .99 as well as β values ranging from .01 to .99. The silhouette coefficient was used as the evaluation criterion. Our findings revealed that the highest silhouette coefficient of 0.782 was achieved using a Cronbach α value of .89, a β value of .97, and 6 topics. Subsequently, we proceeded with topic modeling using the identified parameters.

Analysis of Change in Topics by Period

To address research problem 4—the change in topics in physician competency research over time—the analysis was divided into 4 periods: 2011-2013, 2014-2016, 2017-2019, and 2020-2021. We divided the period into before and after the COVID-19 outbreak, and the researchers discussed and classified them. Subsequently, we analyzed how the percentages of each topic changed. To categorize the topics by checking the pattern of increased or decreased topics by period, we performed a linear regression analysis using SPSS (version 23.0; IBM Corp). We used the categorized periods as independent variables, and the percentage of each topic as the dependent variable. Following the linear regression analysis, we classified the keywords into 4 types based on the regression coefficient sign (+ or −) and the significance probability (P value): hot, warm, cool, and cold. If the coefficient is positive and the significance probability is ≤.05, it is classified as a "hot topic" with increasing research interest. Conversely, if the coefficient is negative and the significance probability is ≤.05, it is classified as a "cold topic" with decreasing research interest. Meanwhile, the clusters that were either positive or negative with no statistical significance and with a significance probability of ≥.05 were classified as "warm" and "cool" topics, respectively [42].

Ethics Approval

This study was conducted after obtaining approval from the Medical Research Ethics Review Committee of Chungnam National University Hospital (CNUH 2021-02-025).

Results

Core Keywords From Physician Competency Studies

Between 2011 and 2020, the words that appeared most frequently in physician competency studies were *knowledge* (604 times), *hospital* (598 times), *family* (597 times), *job* (573 times), *guideline* (491 times), *management* (482 times), and *communication* (443 times). The words that appeared in most studies were *education* (n=256), *model* (n=243), *knowledge* (n=238), and *hospital* (n=234). Table 1 presents the 25 words with the highest frequency and the number of articles in which they appeared.

Table 1. High-ranking keywords by frequency in research.

Rank	Keyword	Frequency, n	Articles in which it appears, n
1	Knowledge	604	238
2	Hospital	598	234
3	Family	597	161
4	Job	573	111
5	Guideline	491	164
6	Management	482	206
7	Communication	443	154
8	Education	437	256
9	Model	425	243
10	Assessment	407	150
11	Attitude	388	151
12	Information	381	213
13	Health care	379	196
14	Experience	368	224
15	Medication	356	64
16	Intervention	332	152
17	Cancer	330	89
18	Disease	313	150
19	Change	276	147
20	Need	268	196
21	Development	255	170
22	Behavior	253	87
23	Barrier	238	97
24	Evidence	234	158
25	Practice	227	199

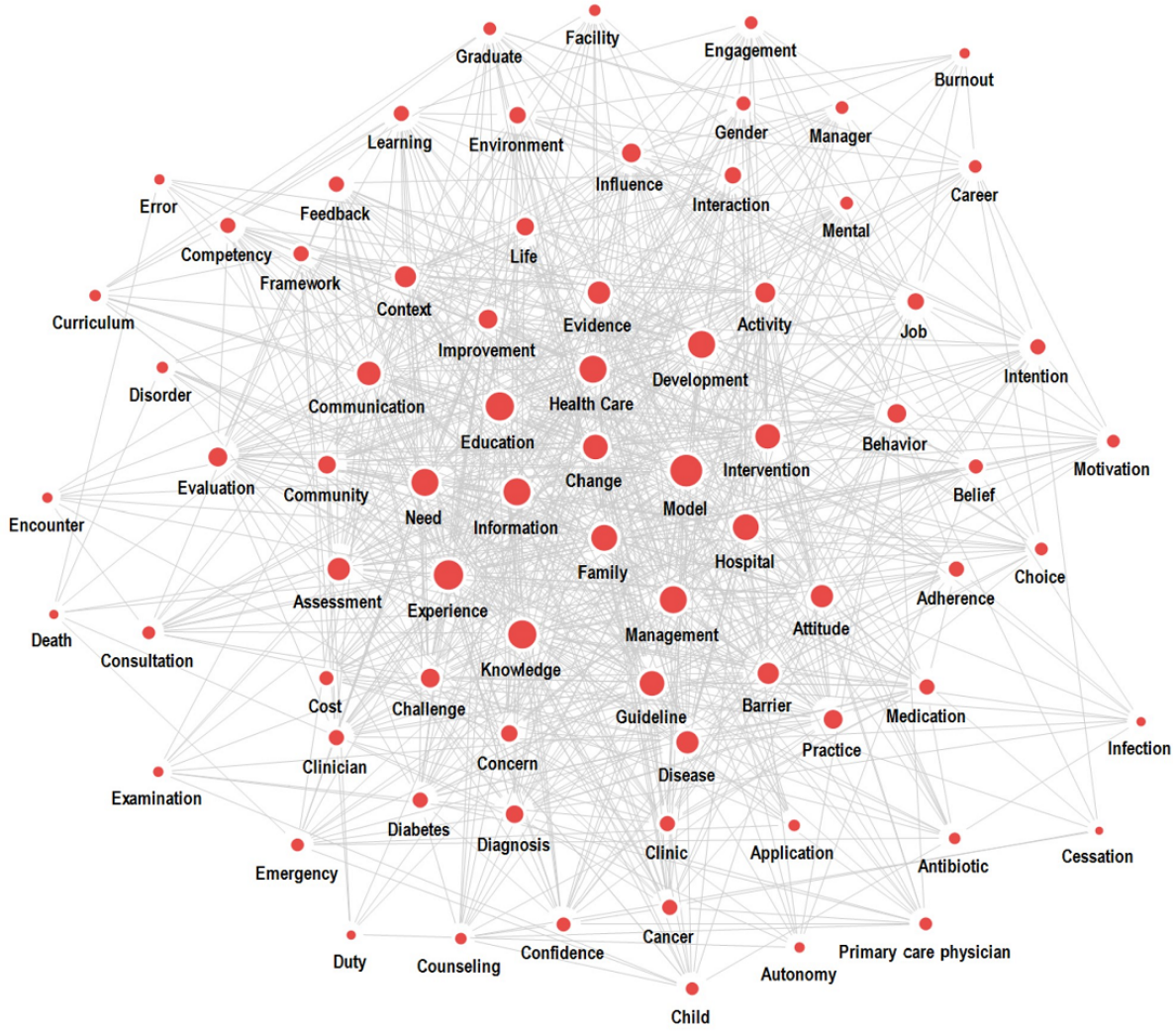
Social Network Analysis

Table 2 presents keyword degree, closeness, and betweenness centrality from physician competency studies. Each keyword and its degree are as follows. The higher the degree of the keyword, the stronger the influence in the network. The keywords with the highest degree and closeness centrality were, in order, *model* (0.988 and 0.988, respectively), *education* (0.981 and 0.982), *experience* (0.973 and 0.975), *health care* (0.973 and 0.975), *hospital* (0.973 and 0.975), and *information* (0.973 and 0.975). The keywords with the highest betweenness centrality were, in order, *model* (0.007), *education* (0.006), *knowledge* (0.006), *management* (0.006), and *education* (0.006). In total, 29 words belonged to the top 30 keywords in all 3 centrality types. The keywords with a high degree centrality were also high in closeness centrality. However, the betweenness centrality of environment was higher than the degree and closeness centrality. Figure 3 shows the network map for centrality.

Table 2. High-ranking keywords by degree, closeness, and betweenness centrality.

Rank	Degree centrality		Closeness centrality		Betweenness centrality	
	Keyword	Degree	Keyword	Degree	Keyword	Degree
1	Model	0.98881	Model	0.98893	Model	0.00651
2	Education	0.98134	Education	0.98169	Education	0.00630
3	Experience	0.97388	Experience	0.97455	Knowledge	0.00616
4	Health care	0.97388	Health care	0.97455	Information	0.00616
5	Hospital	0.97388	Hospital	0.97455	Management	0.00614
6	Information	0.97388	Information	0.97455	Experience	0.00614
7	Knowledge	0.97388	Knowledge	0.97455	Hospital	0.00614
8	Management	0.97388	Management	0.97455	Need	0.00610
9	Need	0.97388	Need	0.97455	Health care	0.00607
10	Development	0.96269	Development	0.96403	Development	0.00600
11	Evidence	0.95522	Evidence	0.95714	Evidence	0.00571
12	Change	0.95149	Change	0.95374	Change	0.00570
13	Intervention	0.94776	Intervention	0.95036	Intervention	0.00551
14	Family	0.93657	Family	0.94035	Disease	0.00543
15	Disease	0.93284	Disease	0.93706	Family	0.00542
16	Guideline	0.92910	Guideline	0.93380	Guideline	0.00512
17	Practice	0.92164	Practice	0.92734	Practice	0.00509
18	Assessment	0.91418	Assessment	0.92096	Assessment	0.00502
19	Challenge	0.91418	Challenge	0.92096	Communication	0.00493
20	Attitude	0.91045	Attitude	0.91781	Challenge	0.00492
21	Communication	0.90672	Communication	0.91468	Attitude	0.00480
22	Evaluation	0.90299	Evaluation	0.91157	Evaluation	0.00474
23	Concern	0.88806	Concern	0.89933	Concern	0.00456
24	Influence	0.88060	Influence	0.89333	Influence	0.00451
25	Activity	0.86940	Activity	0.88449	Improvement	0.00442
26	Improvement	0.86940	Improvement	0.88449	Barrier	0.00430
27	Community	0.86194	Community	0.87869	Activity	0.00427
28	Barrier	0.85821	Barrier	0.87582	Environment	0.00409
29	Implementation	0.85821	Implementation	0.87582	Implementation	0.00408
30	Cost	0.85448	Cost	0.87296	Community	0.00406

Figure 3. Centrality network by the 2D spring network map.



Topic Modeling

Regarding the number of topics from physician competency studies, we decided on 6 topics (silhouette=0.782) by considering the silhouette coefficient and the validity of interpretation. The core keywords by topic are listed in Table 3. The top keywords in topic 1 were *management*, *intervention*, *disease*, *cost*, and *medication*. The top keywords in topic 2 were *family*, *health care*, *information*, *community*, and *need*. The high-ranking keywords in topic 3 were *knowledge*, *attitude*, *cancer*, *guideline*s, and *barrier*s. The high-ranking keywords in topic 4 were, in order, *hospital*, *job*, *burnout*, *model*, *gender*, and *intention*. The top keywords in topic 5 were, in order, *assessment*, *education*, *competency*, *development*, and *graduation*. Finally, the high-ranking keywords in topic 6 were, in order, *communication*, *consultation*, *experience*, *emergency*, and *feedback*.

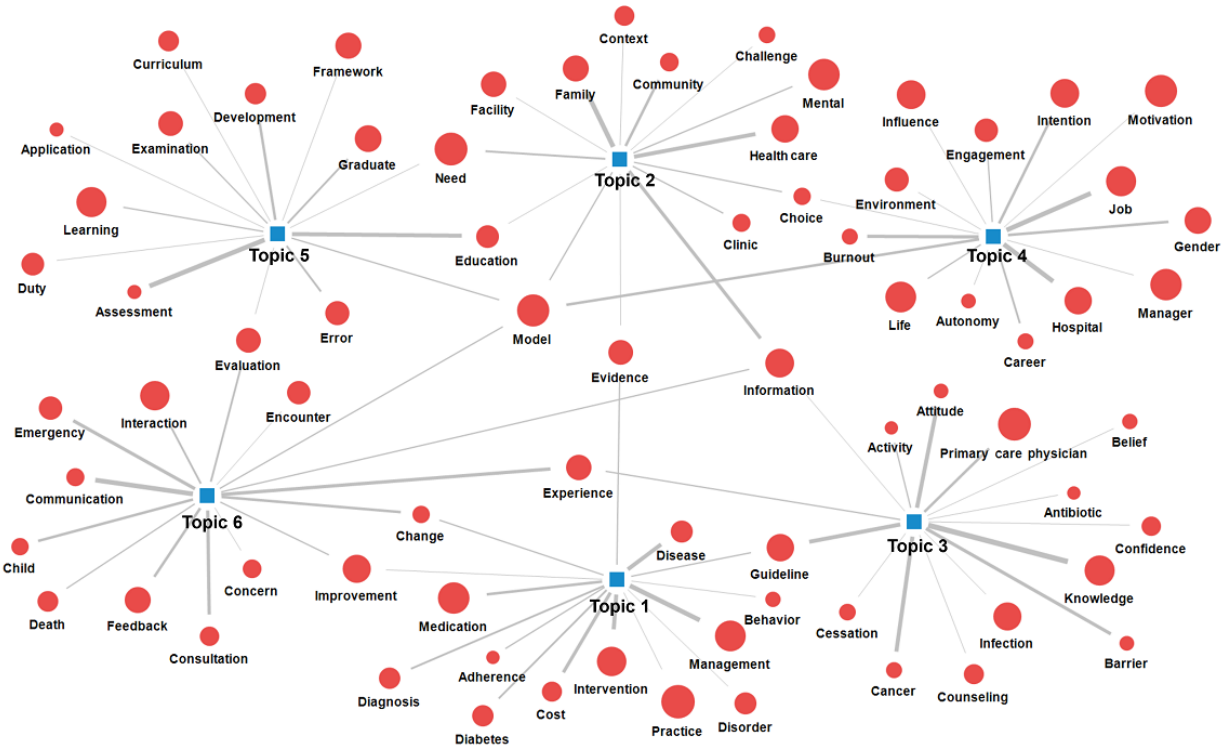
Topic groups were labeled based on high-ranking core keywords in terms of probability distribution: topic 1, *Evidence-based clinical practice*; topic 2, *Community-based healthcare*; topic 3, *Patient care*; topic 4, *Career and self-management*; topic 5, *Continuous professional development*; and topic 6, *Communication and cooperation*. On the basis of the nature of the topics, we divided the topic groups into 2 domains: those related to job competency and those related to personal competency. The job domain includes topics 1, 2, 3, and 6, and the personal domain includes topics 4 and 5. Figure 4 presents the results of visualizing the 7 networks of core keywords using a topic-keyword map.

Table 3. Core keywords by topic.

Rank	Domain and topic					
	Job				Personal	
	Evidence-based clinical practice	Community-based healthcare	Patient care	Communication and cooperation	Career and self-management	Continuous professional development
1	Management	Family	Knowledge	Communication	Hospital	Assessment
2	Intervention	Health care	Attitude	Consultation	Job	Education
3	Disease	Information	Cancer	Experience	Burnout	Competency
4	Cost	Community	Guideline	Emergency	Model	Development
5	Medication	Need	Barrier	Feedback	Gender	Graduate
6	Diagnosis	Choice	PCP[a]	Child	Intention	Error
7	Diabetes	Clinic	Activity	Change	Career	Learning
8	Adherence	Model	Experience	Interaction	Life	Examination
9	Guideline	Mental	Confidence	Evaluation	Engagement	Model
10	Change	Education	Counseling	Model	Manager	Framework

[a]PCP: primary care provider.

Figure 4. Topic modeling network: topic-keyword map.



Core Topics by Period

Table 4 presents the changes in topics over time and Table 5 lists the rate of each topic during this period. Over the past 10 years, each topic has been studied with similar weights. Communication and cooperation accounted for 18.61% (252/1354) of the articles, followed by Evidence-based clinical practice at 17.73% (240/1354) and Continuous professional development at 16.77% (227/1354). Regarding the domains, studies related to physicians' job competencies accounted for 61.21% (910/1354), whereas those related to physicians' personal competencies accounted for 32.79% (444/1354).

Table 4. Number of research on 6 topics by year (n=1354).

Domain and topic	2011 (n=82)	2012 (n=107)	2013 (n=98)	2014 (n=152)	2015 (n=134)	2016 (n=120)	2017 (n=125)	2018 (n=141)	2019 (n=129)	2020 (n=167)	2021 (n=99)	Total (n=1354)
Job, n (%)												
Evidence-based clinical practice	15 (6.3)	24 (10)	24 (10)	27 (11.3)	22 (9.2)	21 (8.8)	25 (10.4)	26 (10.8)	25 (10.4)	21 (8.8)	10 (4.2)	240 (100)
Community-based health-care	11 (5.2)	13 (6.1)	13 (6.1)	22 (10.3)	26 (12.2)	22 (10.3)	17 (8)	18 (8.5)	20 (9.4)	29 (13.6)	22 (10.3)	213 (100)
Patient care	13 (6.3)	22 (10.7)	6 (2.9)	29 (14.2)	12 (5.9)	15 (7.3)	18 (8.8)	23 (11.2)	14 (6.8)	36 (17.6)	17 (17.6)	205 (100)
Communication and cooperation	20 (7.9)	21 (8.3)	22 (8.7)	25 (9.9)	30 (12)	21 (8.3)	22 (8.7)	21 (8.3)	25 (9.9)	24 (9.5)	21 (8.3)	252 (100)
Personal, n (%)												
Career and self-management	11 (5.1)	13 (6)	16 (7.4)	25 (11.5)	19 (8.8)	22 (10.1)	20 (9.2)	24 (11.1)	26 (12)	30 (12)	11 (13.8)	217 (100)
Continuous pro-fessional devel-opment	12 (5.3)	14 (6.2)	17 (7.5)	24 (10.5)	25 (11)	19 (8.4)	23 (10.1)	29 (12.8)	19 (8.4)	27 (12)	18 (7.9)	227 (100)

Table 5. The topic frequency and possession during the period (n=1354).

Domain and topic	Period 1: 2011-2013 (n=287)	Period 2: 2014-2016 (n=406)	Period 3: 2017-2019 (n=395)	Period 4: 2020-2021 (n=266)	Total (n=1354)
Job, n (%)					
Evidence-based clinical practice	63 (22)	70 (17.2)	76 (19.2)	31 (11.6)	240 (17.7)
Community-based healthcare	37 (12.9)	70 (17.2)	55 (13.9)	51 (19.2)	213 (15.7)
Patient care	41 (14.2)	56 (13.8)	55 (13.9)	53 (19.9)	205 (15.1)
Communication and cooperation	63 (225)	76 (18.7)	68 (17.2)	45 (16.9)	252 (18.6)
Personal, n (%)					
Career and self-management	40 (13.9)	66 (16.3)	70 (17.7)	41 (15.4)	217 (16)
Continuous professional develop-ment	43 (15)	68 (16.8)	71 (18)	45 (16.9)	227 (16.8)

Dividing the period into 3-year groups, *Evidence-based clinical practice* (63/240, 26.3%) and *Communication and cooperation* (37/213, 12.9%) were studied most often during the first period (2011-2013), whereas research on *Community-based healthcare* (41/205, 14.3%) and *Career and self-management* (40/217, 13.9%) was conducted on a small scale. *Communication and cooperation* was studied most often during the second period (2014-2016), but the weight decreased compared with the first period (from 22% to 18.7%).

The topics that increased in weight compared with the first period were *Career and self-management* (from 13.9% to 16.3%) and *Continuous professional development* (from 15% to16.8%). During the third period (2017-2019), *Evidence-based clinical practice* (76/240, 19.2%) was studied the most, followed by *Continuous professional development* (71/227, 18%) and *Career and self-management* (70/217, 17.7%). During the fourth period, which covers the COVID-19 pandemic (2020-2021), *Patient care* (53/205, 19.9%) and *Community-based healthcare* (51/213, 19.2%) were the most studied. Compared with the third period, *Patient care* increased from 13.9% to 19.9%, and *Community-based healthcare* increased from 13.9% to 19.2%. Conversely, there were decreases in *Evidence-based clinical practice* (from 19.2% to 11.6%), *Career and self-management* (from 17.7% to 15.4%), and *Continuous professional development* (from 18% to 16.9%).

Regarding the overall possession during the first 3 periods before the COVID-19 pandemic, there was increased research interest in *Career and self-management* (from 13.9% to 16.3% to 17.7%) and *Continuous professional development* (from 15% to 16.8% to 18%) but decreased interest in *Communication and cooperation* (from 22% to 18.7% to 17.2%). In the fourth period, during the COVID-19 pandemic, there was increased research interest in *Community-based healthcare* (45/252, 16.9%) and *Patient care* (53/502, 19.9%) but decreased interest in *Evidence-based clinical practice* (31/240, 11.6%), *Career and self-management* (41/217, 15.4%), and *Continuous professional development* (45/227, 16.9%).

In terms of domains, studies on physicians' personal competencies increased from the first to the third period (from 28.9% to 33% to 35.7%, respectively). However, it decreased

to 32.3% between 2020 and 2021 after the COVID-19 outbreak. Studies on physicians' job competency gradually decreased (from 71.2% to 67.8% to 64.3%), before increasing to 67.8% during the fourth period.

Topic Characteristics by Period

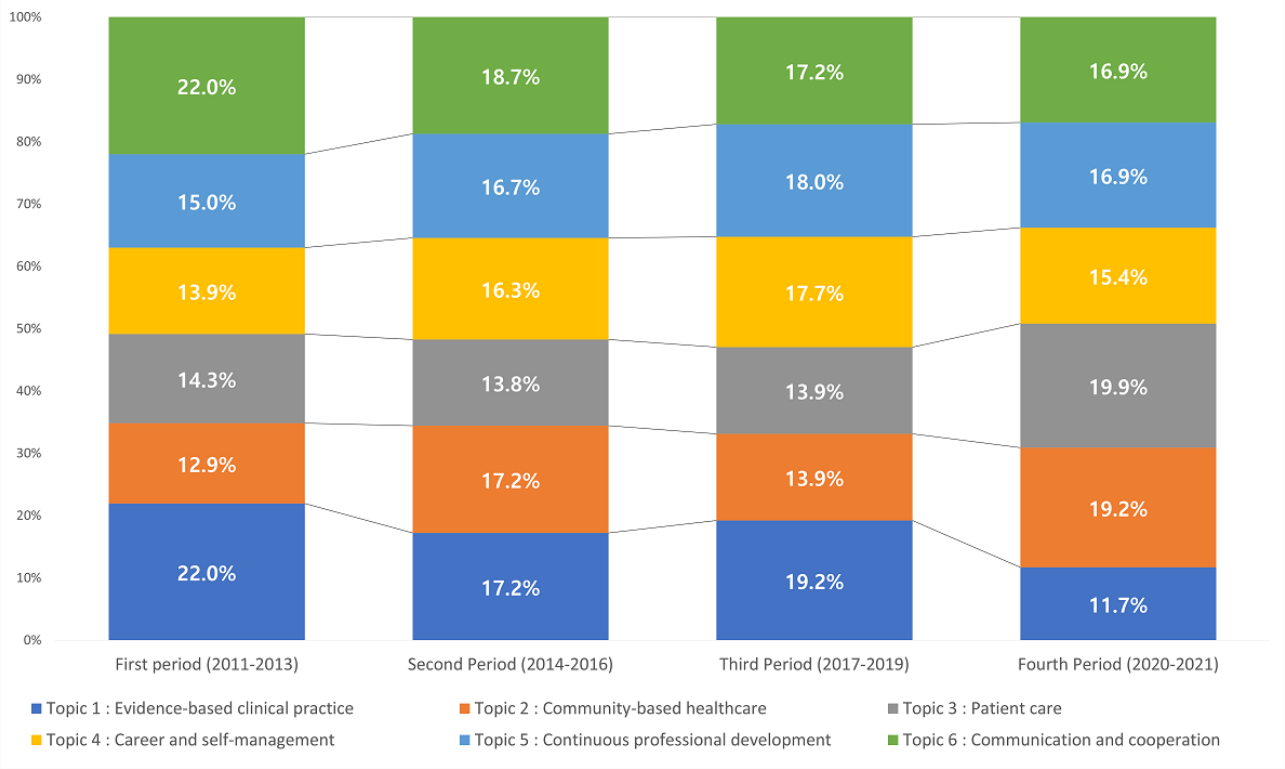
Table 6 presents the topic characteristics for each period. We performed linear regression analysis to examine the characteristics of the 6 topics. Three topics were classified as *hot topics* with a positive regression coefficient and statistical significance: topic 2 (B=1.315; t_9=2.621; *P*=.03), topic 4 (B=1.758; t_9=5.414; *P*=.001), and topic 5 (B=1.339; t_9=2.963; *P*=.02). We found no *cool* (negative regression coefficients and no statistical significance) or *cold* topics (negative regression coefficients and no statistical significance). Figure 5 shows the subject possession during this period.

Table 6. Regression analysis results for each topic.

Domain and topic	B	β	*t* test (*df*)	*P* value	Durbin-Watson statistic	Topic type
Job						
Evidence-based clinical practice	0.388	.339	1.019 (9)	.34	1.535	—[a]
Community-based healthcare	1.315	.680	2.621 (9)	.03	1.181	Hot
Patient care	1.248	.426	1.331 (9)	.22	2.960	—
Communication and cooperation	0.248	.251	0.733 (9)	.48	1.775	—
Personal						
Career and self-management	1.758	.886	5.414 (9)	.001	2.013	Hot
Continuous professional development	1.339	.723	2.963 (9)	.02	2.157	Hot

[a]Not available.

Figure 5. Topic trend during the period.



Discussion

Principal Findings

In this study, we used social network analysis to examine keywords and their relationships in physician competency studies conducted over the past 10 years. Topic modeling identified the top 5 research topics, visualized the relationships among them, and described the research possession over time. Discussions on physicians' competency arose in the 1990s because of the social atmosphere of consumerism, which demands accountability in all aspects of the profession. The medical and health care field emphasizes physicians' roles and attitudes, reflecting the demands of medical consumers, such as citizens [43] and local communities [44]. Educational institutions have received demands to improve the curriculum considering educational outcomes [4,45]. Since the 1990s, some

countries have begun to define the competence of their physicians and specify their components [46]. After the "Project on the future global role of the physician in healthcare" of the World Federation for Medical Education in 2012, it has become more active in many countries [7]. As a goal of the medical community, a competency model for desirable physicians was constructed in the 2000s, centered on Canada [46], the United States [47], the United Kingdom [48], and Scotland [11]. Since then, scholars have actively studied the development of the curriculum and revision of content, reflecting this competency model [8,49,50]. The physician competency studies conducted since 2010, which we analyzed, are on the continuum of physician competency studies carried out over the past 20 years in a large framework.

The words that appeared most frequently in physician competency studies since 2010 are, in order, *knowledge*, *hospital*, *family*, *job*, *guideline*, *management*, and *communication*. The words that appeared in most studies were *education*, *model*, *knowledge*, and *hospital*.

Previously, a physician's professional competence was defined as the habitual and careful use of communication, knowledge, skills, clinical reasoning, emotions, values, and reflection in everyday care for the benefit of individuals and communities [51]. The keywords frequently used in this study had a broad coverage, including the knowledge and context necessary for physicians to perform medical practice, a standardized framework necessary to meet social needs, and traits pertaining to physician groups or individuals.

In the network of top keywords, the keywords with high degree and closeness centrality values were *model*, *education*, *experience*, *health care*, and *hospital*s. Keywords with high betweenness centrality values were *model*, *education*, and *knowledge*. The keyword that emerged in the centrality analysis was *change*. Degree centrality indicates the number of times a node appears simultaneously with other nodes in a network. Closeness centrality connotes the distance between nodes within a network, and betweenness centrality connotes the role of a *bridge* between different nodes within a network. Network centrality analysis of keywords revealed that the 2 central nodes leading physician competency studies over the past 10 years were *model* and *education*. *A model* refers to a pattern, plan, or demonstration that illustrates the structure or work of an object, system, or concept. In competency studies, *model* began with a core competency model that explained what constitutes a physician. Later, scholars focused on different models essential for medical practice, such as specific clinical contexts, resident training, patient-physician communication, leadership [52], and health care management.

Competency-based education for medical students and residents refers to an educational method that uses the content or criteria derived from the previous competency model. In other words, a professional's competency should be built gradually based on scientific knowledge, basic clinical skills, and moral development. Education and experience are essential for acquiring or maintaining expertise in a professional. Song et al [53] conducted a bibliometric analysis of medical expertise from 2010 to 2019. According to these studies, academic journals on medical education primarily include studies on expertise. Likewise, studies on physician competency over the past 10 years have examined methods for developing, implementing, and evaluating competency through education.

Through topic modeling, we identified 6 latent topics. Topic modeling is a researcher-centered content analysis method that identifies a specific pattern assumed to be latent in a document or text and derives a potentially meaningful topic. We also set topic names based on keywords derived from this study. On the basis of relevance of the topic, we divided it into 2 domains: job and personal competency. The job domain comprises *Evidence-based clinical practice*, *Community-based healthcare*, *Patient care*, and *Communicationand cooperation*, whereas the personal domain comprises *Career and self-management* and *Continuous professional development*. The topics derived from topic modeling in this study cover the criteria for physician competency suggested in many countries. In Canada, the Canadian Medical Education Directions for Specialists [46] organized physician roles into 7 competencies: medical experts, communicators, collaborators, leaders, health advocators, scholars, and professionals. In the United Kingdom, "Good Medical Practice 2020" [54] describes physician competency in terms of 4 core competencies: knowledge, skill, and performance; safety and quality; communication, partnership, and teamwork; and trust maintenance. In the United States, the Accreditation Council for Graduate Medical Education laid out 6 core competencies [46] focusing on the areas of physician activities: patient care, professionalism, interpersonal and communication skills, medical knowledge, systems-based practice, and practice-based learning and improvement.

They also comprise the complex knowledge, skills, and attitudes that physicians must possess. Previous studies have emphasized the topics of physician competency research resulting from this study [5]. First, *Evidence-based clinical practice* is a core competency required for clinicians. It provides a framework for integrating research evidence into health care delivery [55], including patient history taking and analysis, physical examination, and diagnostic accuracy. The implementation of evidence-based practice principles has resulted in notable advances in improving the quality of delivered health care [56]. In addition, over the last 20 years, evidence-based practice has been increasingly integrated as a core component of undergraduate, postgraduate, and continuing education health programs worldwide [12]. Second, *Community-based healthcare* is increasingly emphasized in terms of the need for improvement from a broad cultural and institutional perspective to improve the quality of medical care [57]. *Community-based healthcare* has recently attracted more attention, as primary care is emphasized to solve social problems such as population aging and COVID-19 [58].

Third, the topic of *Patient care* means patient-centered care (PCC). PCC enhances health outcomes, such as improved patient satisfaction, behavior change, trust, patient adherence, providers' clinical accuracy, disease management plans, and active patient self-management [59]. Therefore, PCC is a crucial attribute of high-quality health care services [60]. Fourth, physicians' interpersonal and communication skills have a significant impact on patient care. Furthermore, it is correlated with improved

health outcomes and quality [61]. Ineffective communication skills are associated with malpractice claims and suits [62] and medication errors [63]. Communication is a core clinical skill that can be taught and learned [64]. Moreover, interprofessional communication skills are essential competencies for medical students to become physicians [65].

Fifth, among the 2 topics belonging to the personal domain, *Career and self-management* is related to physicians' burnout. Job satisfaction can affect physicians' physical and mental illnesses, such as depression and burnout [66], and is related to patient safety and quality of care [67]. In addition, burnout syndrome is a major concern in occupational health [55]. Therefore, organizations should emphasize the importance of physicians' self-care (rest, a healthy lifestyle, breaks, and sufficient sleep) and regular burnout screening [68]. Finally, *Continuous professional development* refers to the attitude of lifelong learning as a professional. Through professional development such as lifelong learning, medical specialists maintain their professional competence in addition to keeping track of and gaining advancing knowledge [69]. Furthermore, continuing professional development and lifelong learning are crucial for securing high-quality health care, patient safety, and societal trust in the health care system [70]. In other words, the subdomains of physician competency emphasized in many countries have been the main research topics over the past 10 years. Specifically, the main research topics include *Evidence-based medicine* (an explicit means for generating an important answerable question, interpreting new knowledge, and judging how to apply that knowledge in a clinical setting) [71]; *Community-based healthcare* (emphasizes social responsibility); *Patient care* (considering the patient's condition and circumstances throughout the treatment process); *Communication and cooperation* with patients, families, and colleagues; *Career and self-management*; and *Continuous professional development* for maintaining competency as scholars and professionals.

An examination of the changes in research topics over the past 10 years revealed that more studies have been conducted on job competencies than on physicians' personal competencies. Personal domain studies gradually increased from the first to the third period; however, after the COVID-19 outbreak (2020-2021), the number of job domain studies increased.

The most studied topic was *Communication and cooperation*. The topics that showed an increasing frequency and possession before the COVID-19 outbreak (first to third period) were *Career and self-management* and *Continuous professional development*; studies on *Communication and cooperation* showed a decreasing frequency and possession. Studies on *Evidence-based clinical practice* have also gradually decreased (first through second through fourth periods), except during the third period.

Most studies conducted before the COVID-19 outbreak covered physicians' individual professionalism [10,72]. However, during the COVID-19 pandemic (fourth period), studies on *Community-based healthcare* and *Patient care* increased. This can be explained as follows: physicians' social responsibility and community-centered care began to be emphasized during the COVID-19 pandemic, and the importance of care centered on patients and communities has resurfaced. Particularly during the COVID-19 pandemic, many problems threatened patients' health because of the gap in health and medical care, despite individual physicians' expertise and commitment. These challenging social situations highlight the importance of *Community-based healthcare* [73,74].

In times of crisis, the role of physicians can be broadened. For example, physicians have a duty not only to take care of their patients but also to protect them from infection, and thus take care of their families [18]. In addition, social interventions, such as school closures, affect the supply and demand for medical personnel. However, it is not easy to clarify whether a physician's role in situations such as COVID-19 is regular duty. Nevertheless, COVID-19 has broadened the demand for physicians' roles and competencies. This was also manifested in "hot topics" that have gradually increased over the past 10 years, such as *Community-based healthcare*, *Career and self-management*, and *Continuous professional development*. This indicates that topics related to *Community-based healthcare* are gradually becoming more important [73,74], as are topics related to the professionalism of individual physicians [10,51,53,72].

On the basis of the results of this study, the keywords that many researchers were interested in over the past 10 years were *model* and *education*. Therefore, they developed competency-based education and training systems at the hospital, university, and national levels. Consequently, many countries and training institutions, such as hospitals and universities, have developed and educated physicians with competency-based curricula. This was effective in cultivating physicians' competency in responding to social needs.

The topics we should pay attention to are *Community-based healthcare*, *Career and self-management*, and *Continuous professional development*, which are research topics that have gradually increased with time. This indicates that the scope of physicians' competencies has been studied more extensively and comprehensively than in the past. It is meaningful in that it defines physicians' competency as the ability to develop into professional and social leaders, beyond just the ability necessary to perform a job.

The general public's and patients' expectations and consciousness of medical care are changing, and the medical system pursued by society is also changing. Consequently, perceptions of the roles of medicine and physicians are rapidly evolving. In line with these changes, research on the core competencies of physicians must be conducted using more detailed competencies and major fields.

Limitations

This study had some limitations. First, it was difficult to repeat the keyword refining process in the keyword network analysis. To eliminate researcher subjectivity, we have described the analytical procedure in this study. Second, the study period was not equally divided. The last period is short, covering only 6 months, between 2020 and 2021. Because we forecasted that the effect of the COVID-19 pandemic since December 2019

would be reflected between 2020 and the first half of 2021, we set these periods separately. Considering the gap between the time of undertaking and publishing this study, it is unreasonable to argue that the last period accurately reflected the pandemic. However, we believe that it is worth examining the effect of the pandemic because of its global nature, which surpasses its regional and cultural characteristics.

Third, similar to other studies, we explored articles published in English. Physician competency studies are influenced by cultural and social demands and changes in the context of medical services. Although we did not identify the regions in which the studies were published, similar studies on medical professionalism [52] and medical education [75] led us to believe that the English publications in this study were from North America, Europe, or parts of Asia. Future studies should analyze the differences in research trends in physician competency based on culture and region. Finally, this study investigated the core competencies for successfully performing a physician's job, regardless of the specific major. Subsequently, we can research physician's competencies for each major.

Conclusions

The top research topics on physician competency over the past 10 years are *Evidence-based clinical practice*, *Community-based healthcare*, *Patient care*, *Career and self-management*, *Continuous professional development*, and *Communication and cooperation*. The discussion of physician competency entails the establishment of a physician's fundamental roles and competencies based on a constantly changing health care environment and the implementation of education from studying to competency acquisition. Studies on competency include discussions on the model physician desired by society, as well as the issue of wellness encompassing an individual physician's job choice and quality of life.

The hot topics in physician competency studies conducted within the past 10 years are *Community-based healthcare* in the job domain and *Career and self-management* and *Continuous professional development* in the physician's personal domain. These 2 areas are hot topics that have gradually gained interest over time.

Acknowledgments

This paper was supported by Wonkwang University in 2022.

Conflicts of Interest

None declared.

References

1. Brown TC, McCleary LE, Stenchever MA, Poulson Jr AM. A competency-based educational approach to reproductive biology. Am J Obstet Gynecol 1973 Aug 01;116(7):1036-1042. [doi: 10.1016/s0002-9378(16)33856-x] [Medline: 4718215]
2. Spady WG. Competency based education: a bandwagon in search of a definition. Educ Res 1977;6(1):9-14 [FREE Full text] [doi: 10.2307/1175451]
3. Burg FD, Brownlee RC, Wright FH, Levine H, Daeschner CW, Vaughan 3rd VC, et al. A method for defining competency in pediatrics. J Med Educ 1976 Oct;51(10):824-828. [doi: 10.1097/00001888-197610000-00004] [Medline: 972372]
4. Harden RM. AMEE guide no. 14: outcome-based education: part 1-an introduction to outcome-based education. Med Teach 1999;21(1):7-14 [FREE Full text] [doi: 10.1080/01421599979969]
5. Carraccio C, Wolfsthal SD, Englander R, Ferentz K, Martin C. Shifting paradigms: from Flexner to competencies. Acad Med 2002 May;77(5):361-367. [doi: 10.1097/00001888-200205000-00003] [Medline: 12010689]
6. McGaghie WC, Miller GE, Sajid AW, Telder TV. Competency-based curriculum-development in medical-education: an introduction. World Health Organization. 1978. URL: https://apps.who.int/iris/bitstream/handle/10665/39703/WHO_PHP_68.pdf?sequence=1&isAllowed=y [accessed 2021-11-28]
7. Stern DT, Friedman Ben-David M, Norcini J, Wojtczak A, Schwarz MR. Setting school-level outcome standards. Med Educ 2006 Feb;40(2):166-172 [FREE Full text] [doi: 10.1111/j.1365-2929.2005.02374.x] [Medline: 16451245]
8. Tomorrow's doctors: recommendations on undergraduate medical education issued by education committee of the general medical council in pursuance of section 5 of the medical act 1983. General Medical Council. 1983. URL: https://books.google.co.kr/books/about/Tomorrow_s_Doctors.html?id=BDm2xgEACAAJ&redir_esc=y [accessed 2023-06-26]
9. Zaini RG, Bin Abdulrahman KA, Al-Khotani AA, Al-Hayani AM, Al-Alwan IA, Jastaniah SD. Saudi meds: a competence specification for Saudi medical graduates. Med Teach 2011;33(7):582-584. [doi: 10.3109/0142159X.2011.578180] [Medline: 21696288]
10. Ahn D. The future roles of Korean doctors: cultivating well‐rounded doctors. Korean Med Educ Rev 2014 Oct;16(3):119-125 [FREE Full text] [doi: 10.17496/kmer.2014.16.3.119]
11. Simpson JG, Furnace J, Crosby J, Cumming AD, Evans PA, Friedman Ben David M, et al. The Scottish doctor--learning outcomes for the medical undergraduate in Scotland: a foundation for competent and reflective practitioners. Med Teach 2002 Mar;24(2):136-143. [doi: 10.1080/01421590220120713] [Medline: 12098432]
12. Frank JR, Snell L, Sherbino J. CanMEDS 2015 Physician Competency Framework. Ottawa, Canada: Royal College of Physicians and Surgeons of Canada; 2015.

13. Hodges BD, Lingard L. The Question of Competence. New York, NY, USA: Cornell University Press; 2012.
14. Alrassi J, Katsufrakis PJ, Chandran L. Technology can augment, but not replace, critical human skills needed for patient care. Acad Med 2021 Jan 01;96(1):37-43. [doi: 10.1097/ACM.0000000000003733] [Medline: 32910005]
15. Liu XX, Keane PA, Denniston AK. Time to regenerate: the doctor in the age of artificial intelligence. J R Soc Med 2018 Apr;111(4):113-116 [FREE Full text] [doi: 10.1177/0141076818762648] [Medline: 29648509]
16. Longenecker RL, Wendling A, Hollander-Rodriguez J, Bowling J, Schmitz D. Competence revisited in a rural context. Fam Med 2018 Jan;50(1):28-36 [FREE Full text] [doi: 10.22454/FamMed.2018.712527] [Medline: 29346700]
17. Margolies L, Brown CG. Increasing cultural competence with LGBTQ patients. Nursing 2019 Jun;49(6):34-40. [doi: 10.1097/01.NURSE.0000558088.77604.24] [Medline: 31124852]
18. Johnson SB, Butcher F. Doctors during the COVID-19 pandemic: what are their duties and what is owed to them? J Med Ethics 2021 Jan;47(1):12-15 [FREE Full text] [doi: 10.1136/medethics-2020-106266] [Medline: 33060186]
19. Rotenstein LS, Huckman RS, Cassel CK. Making doctors effective managers and leaders: a matter of health and well-being. Acad Med 2021 May 01;96(5):652-654. [doi: 10.1097/ACM.0000000000003887] [Medline: 33332911]
20. Stergiopoulos E, Hodges B, Martimianakis MA. Should wellness be a core competency for physicians? Acad Med 2020 Sep;95(9):1350-1353. [doi: 10.1097/ACM.0000000000003280] [Medline: 32134774]
21. Wiskar K. Physician health: a review of lifestyle behaviors and preventive health care among physicians. BC Med J 2012 Oct;54(8):419-423 [FREE Full text]
22. Dellinger EP, Pellegrini CA, Gallagher TH. The aging physician and the medical profession: a review. JAMA Surg 2017 Oct 01;152(10):967-971. [doi: 10.1001/jamasurg.2017.2342] [Medline: 28724142]
23. Simons MR, Zurynski Y, Cullis J, Morgan MK, Davidson AS. Does evidence-based medicine training improve doctors' knowledge, practice and patient outcomes? A systematic review of the evidence. Med Teach 2019 May;41(5):532-538. [doi: 10.1080/0142159X.2018.1503646] [Medline: 30328793]
24. Cook DA, Oh SY, Pusic MV. Accuracy of physicians' electrocardiogram interpretations: a systematic review and meta-analysis. JAMA Intern Med 2020 Nov 01;180(11):1461-1471 [FREE Full text] [doi: 10.1001/jamainternmed.2020.3989] [Medline: 32986084]
25. Sibeoni J, Bellon-Champel L, Mousty A, Manolios E, Verneuil L, Revah-Levy A. Physicians' perspectives about burnout: a systematic review and metasynthesis. J Gen Intern Med 2019 Aug;34(8):1578-1590 [FREE Full text] [doi: 10.1007/s11606-019-05062-y] [Medline: 31147982]
26. Travaline JM, Ruchinskas R, D'Alonzo Jr GE. Patient-physician communication: why and how. J Am Osteopath Assoc 2005 Jan;105(1):13-18. [doi: 10.7556/jaoa.2005.105.1.13] [Medline: 15710660]
27. Schommer E, Patel VR, Mouraviev V, Thomas C, Thiel DD. Diffusion of robotic technology into urologic practice has led to improved resident physician robotic skills. J Surg Educ 2017;74(1):55-60. [doi: 10.1016/j.jsurg.2016.06.006] [Medline: 27488814]
28. Milner RJ, Gusic ME, Thorndyke LE. Perspective: toward a competency framework for faculty. Acad Med 2011 Oct;86(10):1204-1210. [doi: 10.1097/ACM.0b013e31822bd524] [Medline: 21869668]
29. Saheb T, Saheb M. Analyzing and visualizing knowledge structures of health informatics from 1974 to 2018: a bibliometric and social network analysis. Healthc Inform Res 2019 Apr;25(2):61-72 [FREE Full text] [doi: 10.4258/hir.2019.25.2.61] [Medline: 31131140]
30. Onan A, Korukoglu S, Bulut H. LDA-based topic modelling in text sentiment classification: an empirical analysis. Int J Comput Linguistics Appl 2016;7(1):101-119.
31. Dieng AB, Ruiz FJ, Blei DM. Topic modeling in embedding spaces. Trans Assoc Comput Linguist 2020 Dec;8:439-453. [doi: 10.1162/tacl_a_00325]
32. Jelodar H, Wang Y, Yuan C, Feng X, Jiang X, Li Y, et al. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimed Tools Appl 2019;78(11):15169-15211. [doi: 10.1007/s11042-018-6894-4]
33. Zhao WX, Jiang J, Weng J, He J, Lim EP, Yan H, et al. Comparing twitter and traditional media using topic models. In: Proceedings of the 33rd European Conference on Advances in Information Retrieval. 2011 Presented at: ECIR '11; April 18-21, 2011; Dublin, Ireland p. 338-349 URL: https://link.springer.com/chapter/10.1007/978-3-642-20161-5_34 [doi: 10.1007/978-3-642-20161-5_34]
34. Kim SW, Gil JM. Research paper classification systems based on TF-IDF and LDA schemes. Hum Centric Comput Inf Sci 2019 Aug 26;9(1):30. [doi: 10.1186/s13673-019-0192-7]
35. Kumari R, Jeong JY, Lee BH, Choi KN, Choi K. Topic modelling and social network analysis of publications and patents in humanoid robot technology. J Inf Sci 2021 Oct;47(5):658-676. [doi: 10.1177/0165551519887878]
36. Jonassen DH, Wang S. Acquiring structural knowledge from semantically structured hypertext. J Comput Base Instr 1993;20(1):1-8.
37. Ley T. Knowledge structures for integrating working and learning: a reflection on a decade of learning technology research for workplace learning. Br J Educ Technol 2020 Mar;51(2):331-346. [doi: 10.1111/bjet.12835]
38. Debnath R, Bardhan R. India nudges to contain COVID-19 pandemic: a reactive public policy analysis using machine-learning based topic modelling. PLoS One 2020 Sep 11;15(9):e0238972 [FREE Full text] [doi: 10.1371/journal.pone.0238972] [Medline: 32915899]

39. Song CW, Jung H, Chung K. Development of a medical big-data mining process using topic modeling. Cluster Comput 2019;22(S1):1949-58. Retracted in: Cluster Comput. December 5, 2022. doi: 10.1007/s10586-022-03869-9. [doi: 10.1007/s10586-017-0942-0]
40. O'Callaghan D, Greene D, Carthy J, Cunningham P. An analysis of the coherence of descriptors in topic modeling. Expert Syst Appl 2015 Aug;42(13):5645-5657. [doi: 10.1016/j.eswa.2015.02.055]
41. Mimno D, Wallach H, Talley E, Leenders M, McCallum A. Optimizing semantic coherence in topic models. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. 2011 Presented at: EMNLP '11; July 27-31, 2011; Edinburgh, UK p. 262-272.
42. Griffiths TL, Steyvers M. Finding scientific topics. Proc Natl Acad Sci U S A 2004 Apr 06;101(Suppl 1):5228-5235 [FREE Full text] [doi: 10.1073/pnas.0307752101] [Medline: 14872004]
43. Habbick BF, Leeder SR. Orienting medical education to community need: a review. Med Educ 1996 May;30(3):163-171. [doi: 10.1111/j.1365-2923.1996.tb00738.x] [Medline: 8949549]
44. Hamad B. Community-oriented medical education: what is it? Med Educ 1991 Jan;25(1):16-22. [doi: 10.1111/j.1365-2923.1991.tb00021.x] [Medline: 1997823]
45. Long DM. Competency-based residency training: the next advance in graduate medical education. Acad Med 2000 Dec;75(12):1178-1183. [doi: 10.1097/00001888-200012000-00009] [Medline: 11112714]
46. Frank JR, Danoff D. The CanMEDS initiative: implementing an outcomes-based framework of physician competencies. Med Teach 2007 Sep;29(7):642-647. [doi: 10.1080/01421590701746983] [Medline: 18236250]
47. Swing SR. The ACGME outcome project: retrospective and prospective. Med Teach 2007 Sep;29(7):648-654. [doi: 10.1080/01421590701392903] [Medline: 18236251]
48. Rubin P, Franchi-Christopher D. New edition of tomorrow's doctors. Med Teach 2002 Jul;24(4):368-369. [doi: 10.1080/0142159021000000816] [Medline: 12193317]
49. Cumming A, Ross M. The Tuning Project for Medicine--learning outcomes for undergraduate medical education in Europe. Med Teach 2007 Sep;29(7):636-641. [doi: 10.1080/01421590701721721] [Medline: 18236249]
50. Newble D, Stark P, Bax N, Lawson M. Developing an outcome-focused core curriculum. Med Educ 2005 Jul;39(7):680-687. [doi: 10.1111/j.1365-2929.2005.02198.x] [Medline: 15960788]
51. Epstein RM, Hundert EM. Defining and assessing professional competence. JAMA 2002 Jan 09;287(2):226-235. [doi: 10.1001/jama.287.2.226] [Medline: 11779266]
52. Calhoun JG, Vincent ET, Baker GR, Butler PW, Sinioris ME, Chen SL. Competency identification and modeling in healthcare leadership. J Health Adm Educ 2004;21(4):419-440. [Medline: 15495738]
53. Song X, Jiang N, Li H, Ding N, Wen D. Medical professionalism research characteristics and hotspots: a 10-year bibliometric analysis of publications from 2010 to 2019. Scientometrics 2021;126(9):8009-8027 [FREE Full text] [doi: 10.1007/s11192-021-03993-0] [Medline: 34248230]
54. Good medical practice: protecting patients, guiding doctors. General Medical Council. 2001. URL: https://www.gmc-uk.org/-/media/documents/good-medical-practice-2001-55612679.pdf?la=en%22 [accessed 2023-06-26]
55. De Hert S. Burnout in healthcare workers: prevalence, impact and preventative strategies. Local Reg Anesth 2020 Oct 28;13:171-183 [FREE Full text] [doi: 10.2147/LRA.S240564] [Medline: 33149664]
56. Albarqouni L, Hoffmann T, Straus S, Olsen NR, Young T, Ilic D, et al. Core competencies in evidence-based practice for health professionals: consensus statement based on a systematic review and Delphi survey. JAMA Netw Open 2018 Jun 01;1(2):e180281 [FREE Full text] [doi: 10.1001/jamanetworkopen.2018.0281] [Medline: 30646073]
57. Aveling EL, McCulloch P, Dixon-Woods M. A qualitative study comparing experiences of the surgical safety checklist in hospitals in high-income and low-income countries. BMJ Open 2013 Aug 15;3(8):e003039 [FREE Full text] [doi: 10.1136/bmjopen-2013-003039] [Medline: 23950205]
58. Panigrahi SK, Majumdar S, Galhotra A, Kadle SC, John AS. Community based management of COVID-19 as a way forward for pandemic response. Front Public Health 2021 Jan 13;8:589772 [FREE Full text] [doi: 10.3389/fpubh.2020.589772] [Medline: 33520913]
59. Sultan WI, Sultan MI, Crispim J. Palestinian doctors' views on patient-centered care in hospitals. BMC Health Serv Res 2018 Oct 11;18(1):766 [FREE Full text] [doi: 10.1186/s12913-018-3573-0] [Medline: 30305081]
60. Singh S, Evans N, Williams M, Sezginis N, Baryeh NA. Influences of socio-demographic factors and health utilization factors on patient-centered provider communication. Health Commun 2018 Jul;33(7):917-923. [doi: 10.1080/10410236.2017.1322481] [Medline: 28541816]
61. Nobile C, Drotar D. Research on the quality of parent-provider communication in pediatric care: implications and recommendations. J Dev Behav Pediatr 2003 Aug;24(4):279-290. [doi: 10.1097/00004703-200308000-00010] [Medline: 12915801]
62. Levinson W, Roter DL, Mullooly JP, Dull VT, Frankel RM. Physician-patient communication. The relationship with malpractice claims among primary care physicians and surgeons. JAMA 1997 Feb 19;277(7):553-559. [doi: 10.1001/jama.277.7.553] [Medline: 9032162]
63. Maurette P, Comité analyse et maîtrise du risque de la Sfar. To err is human: building a safer health system. Ann Fr Anesth Reanim 2002 Jun;21(6):453-454. [doi: 10.1016/s0750-7658(02)00670-6] [Medline: 12134587]

64. Rider EA, Keefer CH. Communication skills competencies: definitions and a teaching toolbox. Med Educ 2006 Jul;40(7):624-629. [doi: 10.1111/j.1365-2929.2006.02500.x] [Medline: 16836534]
65. Kim SJ, Kwon OD, Kim KH, Lee JE, Lee SH, Shin JS, et al. Investigating the effects of interprofessional communication education for medical students. Korean J Med Educ 2019 Jun;31(2):135-145 [FREE Full text] [doi: 10.3946/kjme.2019.125] [Medline: 31230436]
66. Kumar S, Fischer J, Robinson E, Hatcher S, Bhagat RN. Burnout and job satisfaction in New Zealand psychiatrists: a national study. Int J Soc Psychiatry 2007 Jul;53(4):306-316. [doi: 10.1177/0020764006074534] [Medline: 17703646]
67. Kravitz RL. Physician job satisfaction as a public health issue. Isr J Health Policy Res 2012 Dec 14;1(1):51 [FREE Full text] [doi: 10.1186/2045-4015-1-51] [Medline: 23241419]
68. Youssef D, Youssef J, Abou-Abbas L, Kawtharani M, Hassan H. Prevalence and correlates of burnout among physicians in a developing country facing multi-layered crises: a cross-sectional study. Sci Rep 2022 Jul 23;12(1):12615 [FREE Full text] [doi: 10.1038/s41598-022-16095-5] [Medline: 35871153]
69. Lombarts KM, Plochg T, Thompson CA, Arah OA, DUQuE Project Consortium. Measuring professionalism in medicine and nursing: results of a European survey. PLoS One 2014 May 21;9(5):e97069 [FREE Full text] [doi: 10.1371/journal.pone.0097069] [Medline: 24849320]
70. van den Goor MM, Wagner CC, Lombarts KM. Poor physician performance in the Netherlands: characteristics, causes, and prevalence. J Patient Saf 2020 Mar;16(1):7-13. [doi: 10.1097/PTS.0000000000000222] [Medline: 26176988]
71. Haynes RB, Sackett DL, Richardson WS, Rosenberg W, Langley GR. How to practice and teach evidence-based medicine. Can Med Assoc J 1997;157(6):788. [doi: 10.1136/bmj.313.7069.1410]
72. Smith R. Medical professionalism: out with the old and in with the new. J R Soc Med 2006 Feb;99(2):48-50 [FREE Full text] [doi: 10.1177/014107680609900202] [Medline: 16449765]
73. Lucey CR, Davis JA, Green MM. We have no choice but to transform: the future of medical education after the COVID-19 pandemic. Acad Med 2022 Mar 01;97(3S):S71-S81 [FREE Full text] [doi: 10.1097/ACM.0000000000004526] [Medline: 34789658]
74. Ohta R, Ryu Y, Sano C. The uncertainty of COVID-19 inducing social fear and pressure on the continuity of rural, community-based medical education: a thematic analysis. Healthcare (Basel) 2021 Feb 17;9(2):223 [FREE Full text] [doi: 10.3390/healthcare9020223] [Medline: 33671392]
75. Ji YA, Nam SJ, Kim HG, Lee J, Lee SK. Research topics and trends in medical education by social network analysis. BMC Med Educ 2018 Sep 24;18(1):222 [FREE Full text] [doi: 10.1186/s12909-018-1323-y] [Medline: 30249248]

Abbreviations

LDA: latent Dirichlet allocation
PCC: patient-centered care

Edited by C Lovis; submitted 06.04.23; peer-reviewed by KH Park, H Hyeonmi; comments to author 02.05.23; revised version received 15.05.23; accepted 16.05.23; published 19.07.23.

Please cite as:
Yune SJ, Kim Y, Lee JW
Data Analysis of Physician Competence Research Trend: Social Network Analysis and Topic Modeling Approach
JMIR Med Inform 2023;11:e47934
URL: https://medinform.jmir.org/2023/1/e47934
doi:10.2196/47934
PMID:37467028

Original Paper

Visual Analytics of Multidimensional Oral Health Surveys: Data Mining Study

Ting Xu[1], PhD; Yuming Ma[2], MSc; Tianya Pan[2], BSc; Yifei Chen[2], BSc; Yuhua Liu[2], PhD; Fudong Zhu[3], PhD; Zhiguang Zhou[2], PhD; Qianming Chen[3], PhD

[1]Department of Stomatology, First Affiliated Hospital, Zhejiang University, Hangzhou, China

[2]School of Media and Design, Hangzhou Dianzi University, Hangzhou, China

[3]The Affiliated Stomatology Hospital Zhejiang University, Hangzhou, China

Corresponding Author:
Zhiguang Zhou, PhD
School of Media and Design, Hangzhou Dianzi University
Xueyuan Road #18
Hangzhou, 310018
China
Phone: 86 15957193211
Email: zhgzhou@hdu.edu.cn

Abstract

Background: Oral health surveys largely facilitate the prevention and treatment of oral diseases as well as the awareness of population health status. As oral health is always surveyed from a variety of perspectives, it is a difficult and complicated task to gain insights from multidimensional oral health surveys.

Objective: We aimed to develop a visualization framework for the visual analytics and deep mining of multidimensional oral health surveys.

Methods: First, diseases and groups were embedded into data portraits based on their multidimensional attributes. Subsequently, group classification and correlation pattern extraction were conducted to explore the correlation features among diseases, behaviors, symptoms, and cognitions. On the basis of the feature mining of diseases, groups, behaviors, and their attributes, a knowledge graph was constructed to reveal semantic information, integrate the graph query function, and describe the features of intrigue to users.

Results: A visualization framework was implemented for the exploration of multidimensional oral health surveys. A series of user-friendly interactions were integrated to propose a visual analysis system that can help users further achieve the regulations of oral health conditions.

Conclusions: A visualization framework is provided in this paper with a set of meaningful user interactions integrated, enabling users to intuitively understand the oral health situation and conduct in-depth data exploration and analysis. Case studies based on real-world data sets demonstrate the effectiveness of our system in the exploration of oral diseases.

(JMIR Med Inform 2023;11:e46275) doi:10.2196/46275

KEYWORDS

visual analytics; oral health data mining; knowledge graph; multidimensional data visualization

Introduction

Background

It is well known that oral health affects systemic health. Oral infections and inflammatory factors have been proven to be highly related to chronic diseases, such as cardiovascular and cerebrovascular diseases and diabetes mellitus [1]. In the field of clinical medicine, oral diseases can be prevented and treated by means of regular professional dental treatments and appropriate oral hygiene practices, which would do great favors for oral health, advance systemic well-being, and improve the quality of life.

As an effective way to investigate oral health status, oral health surveys can determine the frequency, intensity, and spread of oral diseases, such as oral behaviors, oral health cognition, and quality of life in a particular time frame [2]. On the basis of the analysis and mining of oral health surveys, we can obtain deeper

insights into the oral health status of individuals, understand oral diseases, and identify their impacting factors.

However, oral health surveys are always conducted from different perspectives and thus, are presented in the form of multiple dimensions. Traditional data mining methods always use simple statistical charts to visualize surveys, which are limited for the efficient and intuitive mining of deep-seated information. For example, it is difficult to observe the differences and similarities among different oral diseases. The oral health status across different areas and age groups lacks representative descriptions and intuitive comparisons. Thus, it is a difficult task to gain insights from multidimensional oral health surveys, especially for the exploration of relationships among diseases, behaviors, symptoms, and other attributes.

After a series of in-depth discussions with domain experts in the field of stomatology, it was concluded that a visualization system can deliver more comprehensive, interactive, and understandable information to users, which can further help them analyze oral health data rapidly and effectively. However, some challenges remain in implementing an oral health survey–oriented visualization system:

- Challenge 1: both oral diseases and related individuals have different general traits, which makes it difficult to present and compare the different characteristics of oral diseases in addition to their related populations (groups).
- Challenge 2: the occurrence and progression of oral diseases have their own rules, and unhealthy or careless behaviors will prompt the rise of oral diseases. Therefore, it is necessary to investigate the correlations between the diseases and behaviors.
- Challenge 3: oral health surveys include a rich set of data attributes, such as groups, diseases, and behaviors, which present various semantic relationships. It would be of great interest to explore the semantic relations from multidimensional attributes and provide an intelligent retrieval tool based on these relations.

To address the challenges, we developed a visualization framework for the visual analytics and deep mining of multidimensional oral health surveys. First, we designed a set of visualizations to depict the characteristics of diseases and groups combined with multidimensional attributes, such as the struct view, radar view, and cloud view, allowing the comparison of the different traits between various oral diseases and groups (challenge 1). We then designed a scatterplot matrix to analyze the correlation between diseases, behaviors, symptoms, and cognition based on group information, which can further help users discover the relationships among diseases, behaviors, and other attributes (challenge 2). Furthermore, a knowledge graph was designed to integrate diseases, groups, behaviors, and other information, allowing users to gain an overarching view of people with oral diseases. In addition, a query function was provided to conduct personalized retrieval, allowing users to obtain a more detailed understanding of human interests (challenge 3). A visualization framework was implemented to integrate a set of meaningful interactions, allowing users to obtain deeper insights into the patterns of oral diseases according to their requirements. Case studies based on real-world data sets were conducted to demonstrate the effectiveness of our system in visual analytics and deep mining of oral health surveys.

The major contributions of our work are summarized as follows:

- The characteristics of diseases and groups were depicted through portraits in light of multidimensional attributes, enabling users to intuitively and efficiently convey and disseminate information.
- A visualization framework was implemented to enable users to visually analyze and deeply mine the correlation features among oral diseases, behaviors, symptoms, and cognitions.
- A knowledge graph visualization was designed to generate structured knowledge containing semantics, supporting efficient queries on groups or attributes to grasp the semantic characteristics of multidimensional oral health surveys from macro and micro perspectives.

Related Work

This section covers 3 relevant topics: survey data visualization, multidimensional data visualization, and knowledge graph–based data mining.

Survey Data Visualization

Questionnaire survey is a key research tool to uncover and probe the existing states in many research domains [3]. Visualization provides analysts with deeper insights of information through visual recognition. Many researchers have applied data visualizations to realize hidden information capture and personalized exploration of questionnaire data. For example, Drapala et al [4] designed multidimensional data visualizations to explore surveys for the evaluation of information systems. Zhang et al [5] visualized the questionnaire data collected from patients and committed to predetermining orphan disease.

Surveys are widely used in medicine [6]. The World Health Organization provides guidelines for national oral surveys, enabling massive epidemiological studies and discussing survey principles. Powell et al [7] conducted a web-based questionnaire to determine the characteristics of health information seekers visiting a national health service website. O'Brien et al [8] conducted a web-based survey to investigate the use of disability and rehabilitation services among Canadian adults living with HIV. Aggarwal et al [9] piloted a large number of patient samples to explore the patients' views on using their health data in artificial intelligence research. Nakamura et al [10] compared clinicians' and patients' perspectives on treating the symptoms of acute cerebral hemorrhage using survey data.

Multidimensional Data Visualization

Multidimensional data visualization [11] aims to express complex data in a visually intuitive format, using interactive elements to enable users' comprehension of the correlation among various dimensions of the data. With the development of science and technology, multidimensional data have been reflected in a variety of fields. The oral surveys used in this study were multidimensional data with attributes, such as diseases, regions, ages, and behaviors. Currently, multidimensional data visualization includes spatial mapping, glyph [12], small multiples [13], and other methods. Examples

of spatial mapping include scatterplot matrices [14,15], parallel coordinates [16,17], table lenses [18], pixel charts [19], and dimensionality reduction [20,21]. Scatterplot matrices and parallel coordinates are the 2 most widely used multidimensional data visualization strategies. The scatterplot matrix presents high-dimensional data using scatterplots, arranging them based on attributes. This mapping from multidimensional to 2D space helps identify correlations, clusters, outliers, and other notable characteristics. It is a valuable tool for exploring and analyzing complex data sets. Parallel coordinates use a series of parallel axes to represent each variable dimension of high-dimensional data, with the position along each axis corresponding to the variable's value.

In addition to the conventional multidimensional visualizations mentioned above, users can use data portraits [22] to define and describe the various attributes of objects. Data portraits provide a more tangible representation of multidimensional data, allowing for a condensed and effective perception of information panoramas; for example, Xiong and Donath [23] proposed a novel graphical representation based on users' past interactions, encoding people's data with flower and garden metaphors. He et al [24] used accounting indexes to draw the data portrait of the value creation index of all 17 industries, by means of which the characteristics of various industries under COVID-19 can be captured. In this study, we applied data portraits to depict diseases and groups of different regions, genders, and ages.

Knowledge Graph–Based Data Mining

The knowledge graph, introduced by Google in 2012 to refine its search engine, is a typical multilateral relational graph comprising entities and relationships [25]. It serves as a semantic network that reveals the connections between various elements. Knowledge extraction [26], knowledge fusion [27], and knowledge reasoning [28] are the fundamental components involved in constructing a knowledge graph. Knowledge extraction is the process of extracting valuable structured information from large-scale text data, where entity extraction [29] refers to identifying specific entity objects in the text, whereas relation extraction involves extracting the associations and connections between entities. Knowledge fusion involves leveraging technologies such as information extraction, entity alignment, and relationship linking to integrate knowledge from multiple knowledge graphs. This integration results in a more comprehensive, consistent, and accurate knowledge system that enhances knowledge discovery, inference, and application. Knowledge reasoning can generate new factual conclusions by using entity and relation information, thereby expanding the knowledge graph. This process can be categorized into 3 types: logical rule–based reasoning [30], distributed feature representation–based reasoning [31], and deep learning–based reasoning [32].

The data derived from knowledge reasoning can be leveraged in a variety of downstream tasks related to knowledge graphs, such as recommendation systems [33], question answering [34], and information retrieval [35]. For instance, Li et al [36] introduced KG4Vis, a knowledge graph–based visual recommendation method that learns the embedding of knowledge graph entities and relations to capture ideal visual rules. Sousa and Couto [37] provided a new system, named K-BiOnt, by integrating knowledge graphs into biomedical relation extraction, improving the system's ability to identify true relations. Tang et al [38] proposed an intelligent question-answering search system for electric power domain knowledge. The system uses knowledge reasoning to retrieve and analyze information accurately and presents the query results in a visual format. Latif et al [39] developed a visualization system, VisKonnect, to analyze the intertwined lives of historical figures according to the events they participated in through a knowledge graph.

Methods

Oral health data are introduced in this section. A series of analytical tasks are then defined following a thorough discussion with dental experts. Further presentation of the pipeline of our visualization system is encouraged, with the goal of completing the desired analysis tasks.

Data Description

In this study, the real-world data set was obtained from the Oral Health Status Survey and Prevention of Common Diseases in Zhejiang Province [40]. The survey covered several areas, including Jianggan, Hangzhou; Yuyao, Ningbo; Luqiao, Taizhou; Wenling, Taizhou; Wuyi, Jinhua; and Liandu, Lishui. The respondents were from 5 age groups: 3 to 5 years, 12 to 15 years, 35 to 44 years, 55 to 64 years, and 65 to 74 years, representing both urban and rural communities. The data set depicts the oral health status of individuals in 5 age groups in these 6 regions as well as behaviors, symptoms, and cognition associated with oral health. In total, 17 diseases, 14 behaviors, 10 symptoms, and 11 cognitions were considered as research qualities after sorting.

Ethical Considerations

As the data used in the study were deidentified, no ethical approval was sought.

Task Analysis

After detailed discussions with domain experts in the form of structured interviews, we developed a list of analytical tasks for the visual analysis of oral health based on oral health survey reports.

Task 1: How Can the Characteristics of Various Oral Diseases Be Described and Compared?

There are many types of oral diseases, including caries, periodontal disease, and oral mucosal disease. Unfortunately, some simple traditional statistical analyses struggle to uncover the underlying characteristics of the various diseases behind the data. Is a disease, for example, more likely to occur in men or women? At what age group may a malady be more likely to happen? Intuitive and efficient induction will play a vital role in medical research as well as in the formulation of preventive measures.

Task 2: How Can We Describe and Compare the Characteristics of Oral Diseases Among Different Groups?

Different age, region, and gender groups exhibit distinct overall characteristics in terms of prevalence and related attributes. Analyzing various groups based on disease, behavior, symptoms, cognition, and other dimensions can address the limitations of traditional summary evaluations. It enables us to grasp the specific requirements of different groups, aids in the efficient allocation of medical resources to enhance medical service quality, and provides a more comprehensive and precise depiction of the overall disease situation and characteristics of individuals.

Task 3: How Can the Association Among Oral Diseases, Behaviors, Symptoms, and Cognition Be Explored and Presented?

Each disease has its own set of rules regarding its occurrence and development, and it is often the case that bad behavior or carelessness can contribute to the likelihood of developing a disease. Is smoking associated with gum bleeding? Is tooth loss associated with food restriction symptoms? Grasping the relationship between oral diseases, behaviors, cognitions, and symptoms and how these factors interact with each other is a nontrivial task that requires the analysis of intricate and abstract data.

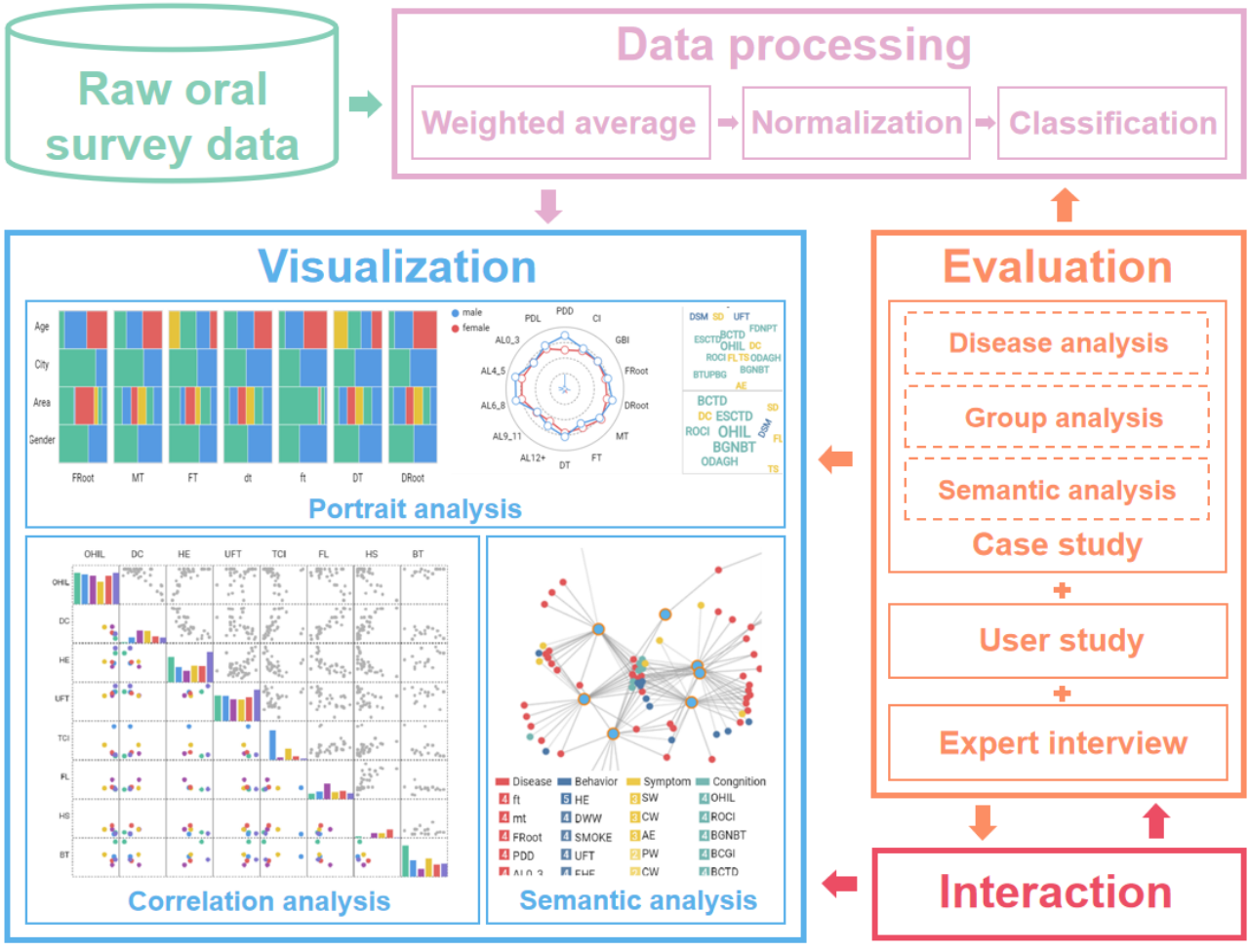
Task 4: How Can Semantic Information of Data Be Revealed in Interpretable Insight and Offer Assistance for Personalized Investigation?

How can we enhance and present the relationship between diseases, behavior, and other factors discovered after mining the correlation? How can users swiftly grasp the traits of specific diseases or groups at a detailed level? In addition, how can users gain a comprehensive understanding of the overall semantic context encompassing diseases, groups, behaviors, and other attributes from a macro perspective? Using effective visualizations can substantially aid users in comprehending and preventing oral diseases, promoting health awareness, and fostering healthy coping strategies.

System Overview

In this study, we developed a multidimensional survey visualization system for oral health that enables users to perceive the characteristics and patterns of oral diseases. Figure 1 shows the pipeline of the system to illustrate the design and implementation of the visualization framework.

Figure 1. The pipeline of our visualization system.



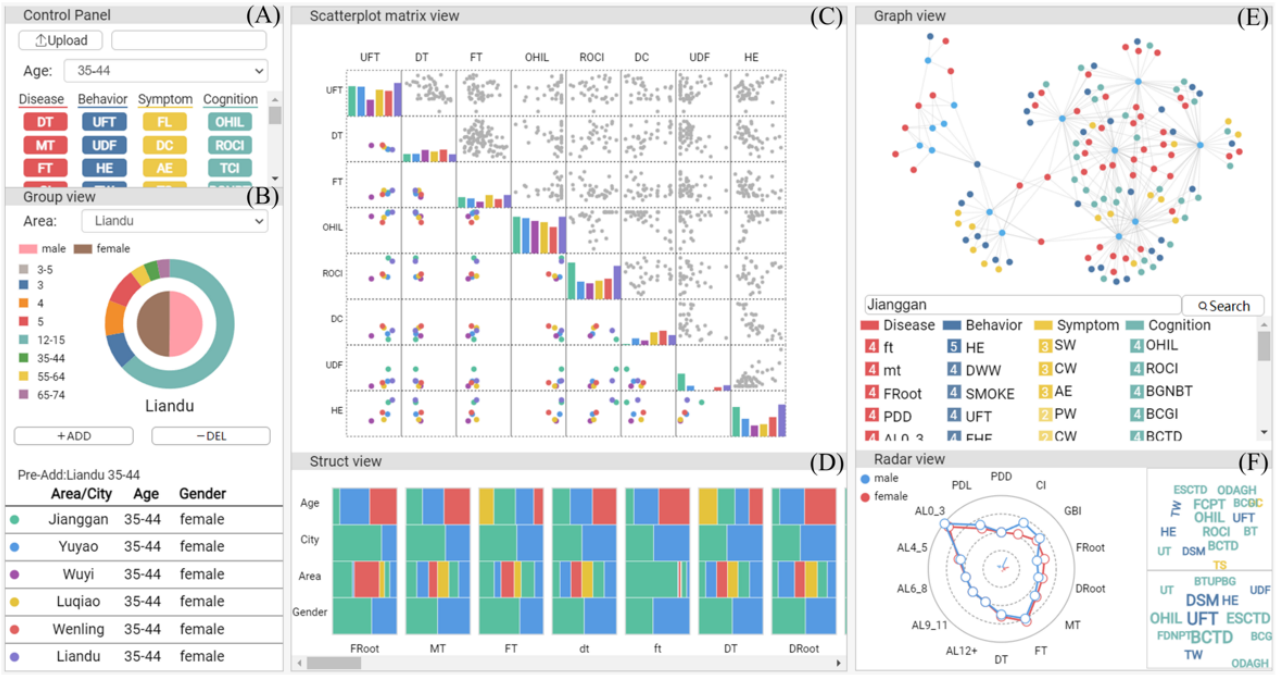
Before visualization, the original oral survey report data set is loaded and preprocessed in 3 steps: weighted evaluation, normalization, and classification. Then, rich visualizations are leveraged for the visual analysis of multidimensional oral health

data, focusing on portrait analysis, correlation analysis, and semantic analysis. Various types of views, including the struct view, radar view, and cloud view, are used to depict the data portraits of diseases and groups. A scatterplot matrix view can be used to examine the correlation of attributes between groups and analyze the differences between groups. A graph view is used to link groups, diseases, behaviors, symptoms, and cognitions, thereby uncovering the semantic associations. Thereafter, the system's effectiveness in revealing multidimensional oral health surveys is evaluated through case studies, user studies, and expert interviews. In addition, a rich visual interface and user-friendly interactions are provided for users to explore the multidimensional oral health data in depth.

Visual Exploration

We developed a user-friendly visualization system to help users observe oral disease characteristics, disease-behavior correlations, and the semantic information of diseases and multidimensional attributes. This system includes a control panel (Figure 2A), group view (Figure 2B), scatterplot matrix view (Figure 2C), struct view (Figure 2D), graph view (Figure 2E), and radar view (Figure 2F), offering user-friendly interaction.

Figure 2. The visualization interface for a multidimensional oral health survey. (A) A control panel enabling users to load data sets and select attribute labels for different groups. (B) The group view to show the composition features and allow users to select groups. (C) The scatterplot matrix view to reveal the correlation features of attributes. (D) The struct view to present the characteristics of oral diseases. (E) The graph view illustrating the semantic knowledge of attributes such as population and disease. (F) The radar view with 2 cloud views to reveal the characteristics of different groups on attributes such as diseases and behaviors. AE: ashamed to eat; AL: attachment loss; BCGI: Bacteria can cause gum inflammation; BCTD: Bacteria can cause tooth decay; BGNBT: Bleeding gums are normal when brushing teeth; BT: brush the teeth; BTUPBG: Brushing teeth is useless in preventing bleeding gums; DC: difficulty chewing; CW: communication worry; DRoot: decayed teeth root due to caries; DROOT: decayed teeth root due to caries; DT: decayed tooth; DW: dietary worry; ESCTD: Eating sugar can cause tooth decay; FCPT: Fossa closure can protect the teeth; FDNPT: Fluoride does not protect teeth; FHE: father's highest education; FL: food limitation; FRoot: fill teeth root due to caries; FROOT: fill teeth root due to caries; ft: fill deciduous tooth due to caries; FT: fill tooth due to caries; GBI: bleeding gums index; HE: highest education; HS: hinder to speak; MHE: mother's highest education; MT: missing tooth; OC: only child; OHIL: Oral health is important to life; PDD: periodontal pocket depth; PDL: periodontal pocket depth 4~6mm; PW: pronunciation worry; ROCI: Regular oral check-ups are important; SD: swallowing discomfort; ST: Time since the last dental visit; SW: sleep worry; TCI: Tooth condition is innate, not acquired; TS: tooth sensitivity; TW: toothwash within 12 months; UT: Use the toothpick.



Data Portrait Analysis

Struct View

To visually and effectively depict different diseases, we designed the struct view. Disease profiles are presented as large rectangles containing smaller rectangles, categorized by area, gender, age, and city. Each small rectangle's width represents the prevalence rate of the disease, whereas its color is randomly chosen. Clicking a small rectangle on the screen will display the region, gender, age, and urban and rural areas, along with the prevalence value. In total, 17 diseases were identified. To accommodate the limited screen space and enhance visual presentation, users can use sliding blocks.

Radar View

Radar view is a widely used metaphor in visualization. It allows for the presentation and comparison of group characteristics based on area, age, and gender. We developed a radar view specifically for this purpose. When the user selects a group, the radar view displays the disease attribute values in the specified region and age group for both men and women. Each axis maps the number of teeth with a type of oral disease to comprehend and compare the oral disease status of men and women in a fixed region and fixed age from a single radar view. By choosing different radar views, we can also compare different age groups and regional groups from the list.

Cloud View

To indicate distinct behaviors, symptoms, and cognitions of each group and explore their correlation with diseases, we offer cloud views that correspond to male and female groups in addition to the radar view. The original data provide the prevalence rate and average number of teeth for the disease attribute in each group. However, the attributes related to behaviors, cognition, and symptoms are usually represented by the proportions of individuals in different degrees. Drink sweat milk, for instance, divides individuals into 4 categories: seldom, 1 per month, 1 per week, and ≥1 per day. It is challenging to evaluate and compare the strengths of each group in these attributes. To address this, we calculated the weighted average and converted it to a value between 0 and 1, known as the strength value. Words are color-coded based on their attribute category, and their size indicates the attribute strength. In this way, users can gain a more intuitive understanding of the behaviors, cognition, and symptoms of different groups and compare them to disease features in the radar view.

Correlation Analysis

We used a scatterplot matrix (Figure 2C) to infer the correlations among complex data attributes. The scatterplot matrix, an extension of the scatterplot for multidimensional data, is crucial for visualizing binary relationships. Nevertheless, the number of matrix elements that can be displayed is constrained by screen size when there are too many dimensions. Here, we applied internet-based methods for users to select the disease, behavior, cognition, and symptom tags in the control panel. The selected tags served as dimensions in a scatterplot matrix view. As we need to use groups to model the associations between the attributes, we first selected several groups within the group view (Figure 2B). In the scatterplot matrix view located on the diagonal, the histogram is used to show and compare the performance of the selected groups in the corresponding attributes. For disease attributes, the height of the bar maps the prevalence, whereas for behaviors, cognitions, or symptoms, the height of the bar depicts the strength value. Scatterplots outside the diagonal are deployed to present 2-by-2 relationships between attributes, with each scatter representing a group. The scatterplot has 2 dimensions: the strength value of the attribute or the normalized number of teeth. Owing to their symmetrical nature, we designed a matrix above the diagonal that offers relationships for all groups, whereas the matrix below the diagonal presents the relationships of the currently selected group under the corresponding attributes. The distribution of scattered points serves as a visual representation of the correlation between the multidimensional attributes.

Semantic Analysis

How can acquired knowledge be logically and scientifically presented after obtaining relevant features and information? We constructed a large-scale knowledge graph (Figure 2E) consisting of entities such as groups, diseases, behaviors, symptoms, and cognitions and established relations among groups and diseases, behaviors, symptoms, and cognitions. These relations can be broken down into 4 categories: *group has diseases*, *group holds behaviors*, *group shows symptoms*, and *group carries cognitions*. There are 112 group entities in the graph; for example, “Jianggan District, 12-15 years old, female.”

To characterize groups based on multidimensional attributes and establish the association between groups, we categorized the average tooth value for disease and the strength values of behavior, symptoms, and cognition under the guidance of experts. By considering the numerical distribution of all groups across each attribute and their respective strengths, we divided them into 4 to 5 categories. There are 66 disease entities, including the 4 categories of decayed tooth (DT) due to caries: *almost noDT (DT1)*, *mild DT (DT2)*, *moderate DT (DT3)*, and *severe DT (DT4)*. Similarly, behaviors, symptoms, and cognitions were classified into entities according to their strength value, with a total of 38 behavior entities, 30 symptom entities, and 60 cognition entities. This enables semantic associations among groups, diseases, behaviors, symptoms, and cognitions through the knowledge graph, providing users with a precise and comprehensive semantic expression. To efficiently extract group or attribute characteristics from large-scale entities, we set up a search function in our hub. Users can input keywords related to the node they wish to query, such as groups, diseases, behaviors, cognitions, and symptoms. When searching for a single group or multiple groups, associated disease, behavior, cognition, and symptom entities will be displayed below the search box, organized by category, and ranked to provide a visual description of various attributes of the group. Users can infer the risk of oral diseases through their own similar groups from group-related behaviors, cognitions, symptoms, and other attributes and further grant decision evidence for the prevention of oral diseases. Thus, the visualization tool can easily allow users to identify and intervene in potential oral disease risks and enable medical teams to formulate personalized prevention strategies.

Visual Interface

We provided a rich set of interactions to assist users in conducting an in-depth analysis of multidimensional oral health surveys. Groups, diseases, behaviors, cognitions, and symptoms can be selected by users, thus enabling personalized and targeted exploration. We offer operations for data loading in the control panel, as shown in Figure 2A. Users can select labels from 4 categories on the control panel: disease, behavior, symptom, and cognition labels. We provided different labels for different age groups because of substantial variations in survey data across age groups. We provided a nested pie chart to display the sample size composition of the groups in 6 regions, including urban and rural areas. The inner circles indicate gender, whereas the outer circles indicate age. Users can select a region, choose age and gender within the corresponding pie chart, and click the “Add” button to include the selected group. After adding the groups, they are displayed in a list below the button, distinguished by random colors. In this way, the chosen label determines the attributes of the scatter matrix, whereas the chosen groups facilitate attribute comparison in the diagonal. All attribute features of the chosen group are also visible in the cloud view and radar view. The graph view shows the semantic relationships of all groups and attributes and offers search capabilities to aid in investigating specific groups and attributes.

Results

Overview

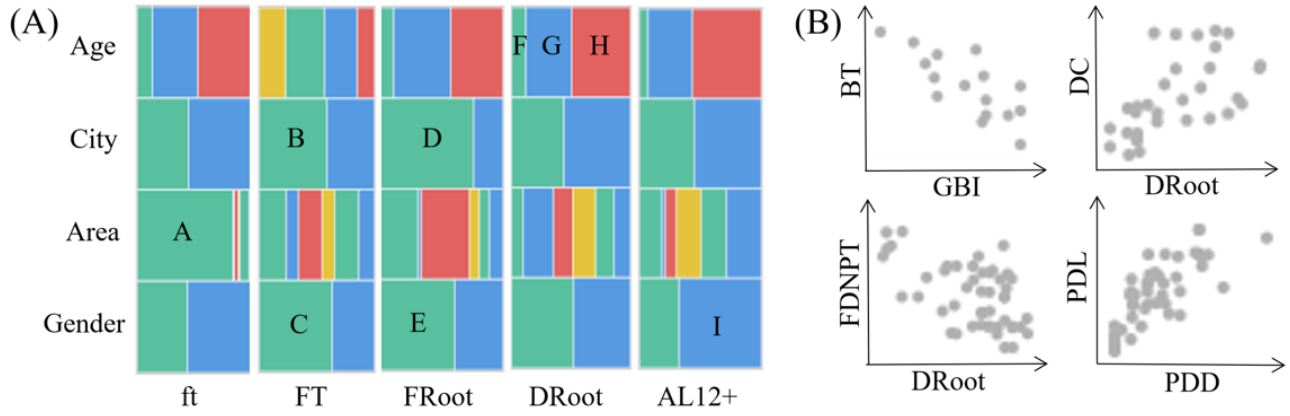
As a web-based visual analysis system, this system was developed using the classic front-end–based frame of ES6+d3.js+csv. A Windows platform with a 2.3 GHz Intel Core i7 CPU and 16 GB of memory was used as the front-end page server. The evaluation experiments were performed using a Google Chrome web browser. Our system can facilitate the efficient and intuitive information mining of experts and users regarding oral diseases. Case studies, user studies, and expert interviews were conducted to demonstrate the usability and viability of our system.

Case Study

Case 1: Disease Analysis

Each disease has its own characteristics, and we introduced a structural view to reveal the characteristics of various diseases. As shown in Figure 3A, we recorded the patient composition of 5 diseases, fill deciduous tooth due to caries (ft), fill tooth due to caries (FT), decayed teeth root due to caries (DRoot), fill teeth root due to caries (FRoot), and attachment loss (AL)12+, in the 4 dimensions of age, city (urban or rural), area, and gender. We found that the width of block A is much larger than that of the 5 blocks on its right. Block A represents the proportion of people with ft in Jianggan. We examined the economic situation of 6 regions with this question in mind. Jianggan exhibits the best overall development among the 6 areas, which explains the higher prevalence of children with filling caries in Jianggan. Blocks B, C, D, and E represent the proportion of urban and female patients with FT and FRoot, respectively. These findings suggest that urban residents possess a better understanding of filling decayed teeth and roots compared with rural residents, and women demonstrate higher awareness than men. Blocks F, G, and H illustrate the proportions of individuals, aged 35 to 44 years, 55 to 64 years, and 65 to 74 years, with DRoot, which shows that the risk of DRoot will increase with age. Similarly, AL12+ is also more likely to occur in the older adult population. Block 1 represents the proportion of men with AL12+, which indicates that men are more likely to experience significant periodontal AL.

Figure 3. Characteristics and risk factors for diseases. (A) Portraits selected from the struct view. (B) Attribute correlations selected from scatterplot matrix. AL12+: attachment loss ≥12 mm; BT: brush the teeth; DC: difficulty chewing; DRoot: decayed teeth root due to caries; FDNPT: fluoride does not protect teeth; ft: fill deciduous tooth due to caries; FT: fill tooth due to caries; FRoot: fill teeth root due to caries; GBI: bleeding gums index; PDL: periodontal pocket depth 4~6mm; PDD: periodontal pocket depth ≥6 mm.



"You get what you grow, you get what you grow." Oral diseases do not just appear out of nowhere; they are frequently tightly tied to certain actions and cognition. These diseases bring symptoms that affect our daily lives, and some can even spread and lead to other illnesses. As shown in Figure 3B, we intercepted 4 pairs of examples of correlation from the scatterplot matrix view: disease and behavior, disease and symptom, disease and cognition, and disease and disease. We can see that bleeding gums index (GBI) is negatively correlated with brush the teeth, DRoot is positively correlated with difficulty chewing, DRoot is negatively correlated with fluoride that does not protect teeth, and periodontal pocket depth (PPD) of ≥6 mm was positively correlated with the periodontal pocket length of 4-6 mm.

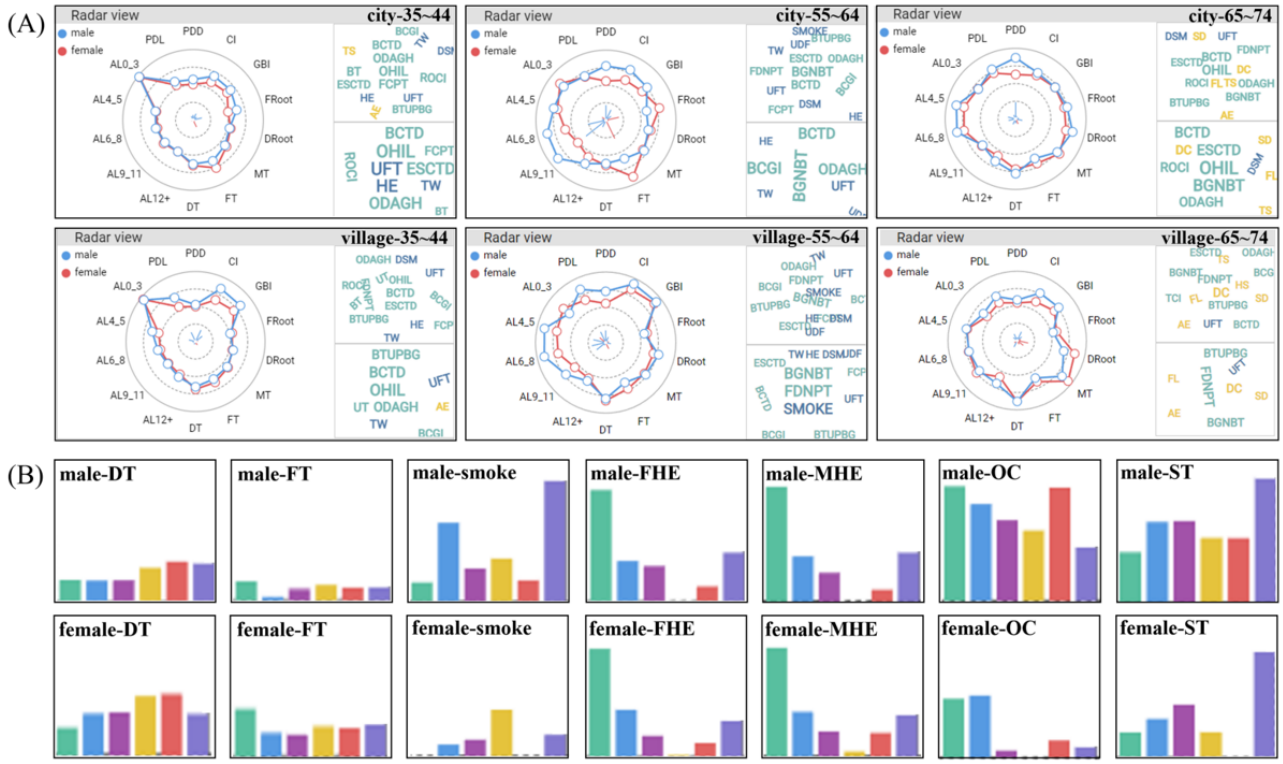
We have summarized some information after a thorough disease analysis. The periodontal health of men is significantly lower than that of women, and individuals lack caries awareness and treatment, and the rate of caries filling treatment is generally low. Middle-aged and older adult groups need to take more proactive measures to prevent and treat periodontal disease in rural areas, which is significantly lower than that in urban areas.

Case 2: Group Analysis

We proceeded with our investigation of the group after examining the characteristics of diseases. The radar view and cloud view in Figure 4A show the diseases, behaviors, symptoms, and cognitions of urban and rural groups of people aged 35 to 44 years, 55 to 64 years, and 65 to 74 years. The horizontal comparison reveals age characteristics, the vertical comparison reveals urban and rural characteristics, and the figure reveals gender characteristics. We learned that the overall disease conditions of the 35 to 44 years age group were less severe than those of the older group. With increasing age, AL, missing tooth (MT) due to caries, DRoot, and DT due to caries showed a deteriorating trend. In the cloud view, the number of yellow words representing symptoms of the group from 65 to 74 years was significantly higher than that of the age group from 35 to 34 years (data for the 55-64 years age group are unavailable and thus not included). When comparing urban and

rural regions, it is evident that the occurrence rates of calculus index, GBI, and DT due to caries were higher in rural areas, whereas FRoot and FT exhibited lower rates. Behaviors such as highest education, use fluoride toothpaste, and toothwash within 12 months, indicated in blue, and cognition, indicated in green, were generally stronger in cities than in villages. Analyzing men and women (55-64 years age group) in the city through the radar view and cloud view, it is apparent that men have greater severity of periodontal diseases, such as PPD ≥6mm, calculus index, GBI, and AL. On the other hand, women showed significantly higher occurrences of FRoot and FT than men, suggesting that women in this group were more conscious of filling teeth due to caries.

Figure 4. Characteristics and comparison of groups. (A) Radar views and cloud views of male and female groups of 3 ages in city and village. (B) The bar charts in the scatterplot matrix of boy and girls aged 12-15 years in each attribute in 6 regions. AE: ashamed to eat; AL: attachment loss; BCGI: Bacteria can cause gum inflammation; BCTD: Bacteria can cause tooth decay; BGNBT: Bleeding gums are normal when brushing teeth; BTUPBG: Brushing teeth is useless in preventing bleeding gums; DC: difficulty chewing; DRoot: decayed teeth root due to caries; DROOT: decayed teeth root due to caries; DT: decayed tooth; ESCTD: Eating sugar can cause tooth decay; FCPT: Fossa closure can protect the teeth; FDNPT: Fluoride does not protect teeth; FHE: father's highest education; FL: food limitation; FROOT: fill teeth root due to caries; ft: fill deciduous tooth due to caries; FT: fill tooth due to caries; GBI: bleeding gums index; HE: highest education; MHE :mother's highest education; MT: missing tooth; OC: only child; OHIL: Oral health is important to life; PDD: periodontal pocket depth; PDL: periodontal pocket depth 4~6mm; ROCI: Regular oral check-ups are important; ST: time since the last dental visit; TCI: Tooth condition is innate, not acquired; TS: tooth sensitivity.



The scatterplot matrix views facilitate intuitive comparison of attributes between different groups, as depicted in Figure 4B. This figure presents behavior comparisons between boys and girls aged 12-15 years in 6 regions: Jianggan, Yuyao, Luqiao, Wenling, Wuyi, and Liandu (from left to right). Each column represents a region, allowing for comparisons across different regions within each subpart. For example, the attributes of father's highest education and mother's highest education indicate that boys and girls in Jianggan outperform those in other regions. In addition, Liandu exhibits a longer time since the last dental visit (ST) compared with others. In the attributes smoke, only child, and ST, boys are significantly more numerous than girls, while in DT and FT, girls are more numerous than boys.

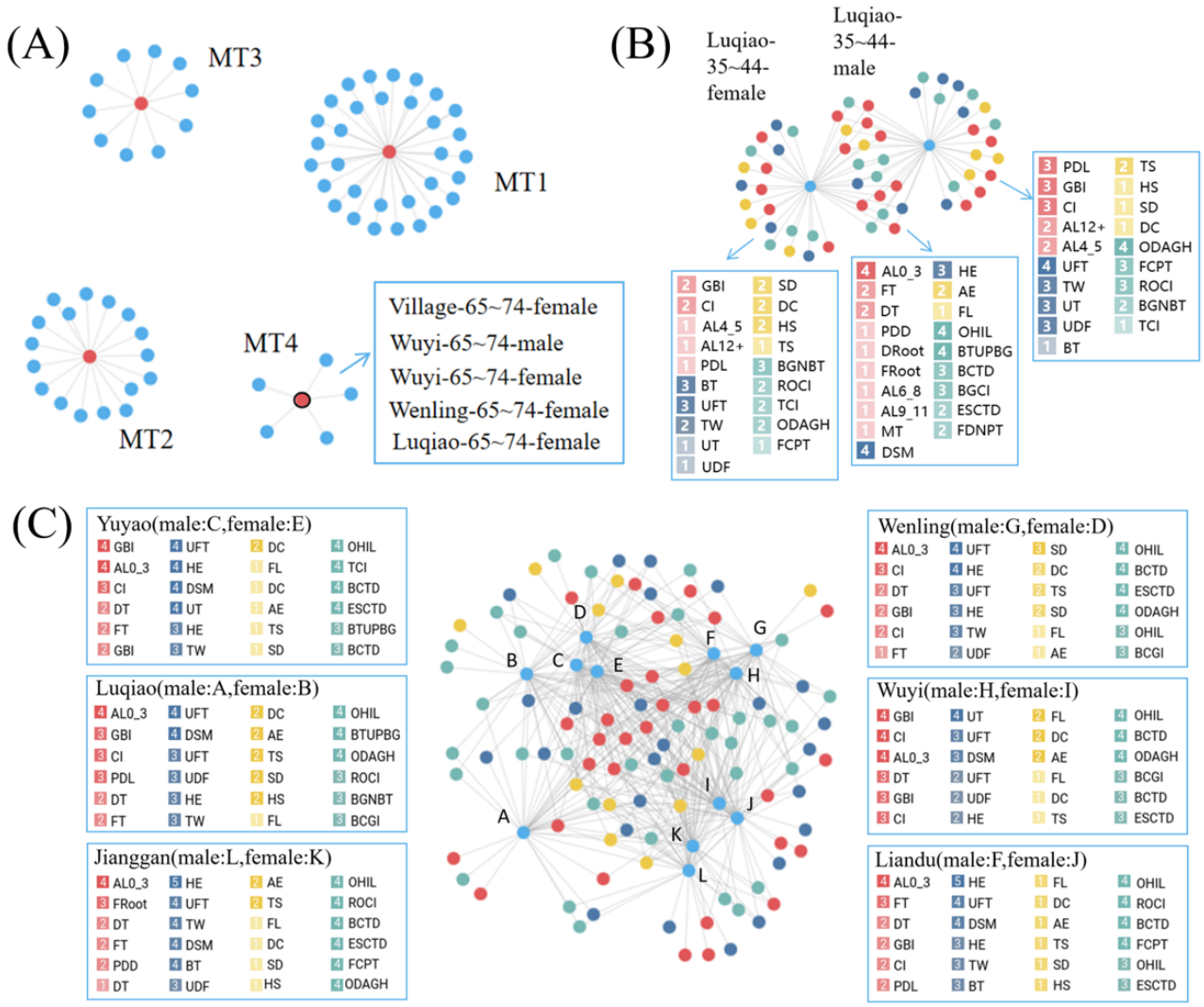
We summarized some information after a comprehensive group analysis. There are differences in oral health conditions. The higher quality of dental care and periodontal health in rural areas compared with urban areas may be because of the difference in economic prosperity and level of education and knowledge about oral health between the two. There are gender differences in oral health conditions. The mean and rate of caries in women were slightly higher than those in men, whereas the number of caries fillings and periodontal health in women were better than those in men. There are age differences in the oral health conditions. The prevalence of dental loss and periodontal diseases increased as individuals aged, especially among the older adult group with a weak awareness of the treatment of dental loss and caries. This information can aid medical teams in developing targeted and personalized prevention strategies to address these gender and age disparities in oral health.

Case 3: Semantic Analysis

We constructed a macro knowledge graph for groups, diseases, behaviors, symptoms, and cognitions, effectively converting complex and diverse objects into accessible and intuitive information. Figure 5A shows the semantic association between the different degrees of MT due to caries and the population.

MT1, MT2, MT3, and MT4 represent disease severity, with the population classified accordingly. Not only diseases but also behaviors, symptoms, and cognitive attributes can be categorized. Figure 5A displays the information of the group linked to the MT4 entity node, revealing that the severe dental disease group consisted entirely of older adults aged 65 to 74 years.

Figure 5. Discovery of query based on knowledge graph. (A) Similarity and difference attribute information of different degrees of MT due to caries and population. (B) Similarity and difference attribute information of different gender groups. (C) Similar and difference attribute information of different area groups. AE: ashamed to eat; AL: attachment loss; BCGI: Bacteria can cause gum inflammation; BCTD: Bacteria can cause tooth decay; BGNBT: Bleeding gums are normal when brushing teeth; BT: brush the teeth; BTUPBG: Brushing teeth is useless in preventing bleeding gums; CW: communication worry; DC: difficulty chewing; DRoot: decayed teeth root due to caries; DROOT: decayed teeth root due to caries; DT: decayed tooth; DW: dietary worry; ESCTD: Eating sugar can cause tooth decay; FCPT: Fossa closure can protect the teeth; FDNPT: Fluoride does not protect teeth; FHE: father's highest education; FL: food limitation; FROOT: fill teeth root due to caries; ft: fill deciduous tooth due to caries; FT: fill tooth due to caries; GBI: bleeding gums index; HE: highest education; HS: hinder to speak; MHE: mother's highest education; MT: missing tooth; OC: only child; OHIL: Oral health is important to life; PDD: periodontal pocket depth; PDL: periodontal pocket depth 4~6mm; PW: pronunciation worry; ROCI: Regular oral check-ups are important; SD: swallowing discomfort; ST: Time since the last dental visit; SW: sleep worry; TCI: Tooth condition is innate, not acquired; TS: tooth sensitivity; TW: toothwash within 12 months; UT: Use the toothpick.



After exploring the groups corresponding to the attributes, we further explored the attributes of groups. Figure 5B illustrates the corresponding attribute association between men and women aged 35 to 44 years in Luqiao, and the commonalities and differences between groups are intuitively presented. For example, AL in both groups was mild, whereas men were more likely than women to have PDL.

Figure 5C illustrates the data for those aged 35 to 44 years in 6 regions. It provides a comprehensive overview of the semantic relationships between these groups, highlighting shared characteristics as well as unique attributes. Notably, we observed a consistent association between the groups facilitated by distinct attributes. Groups from the same region or gender exhibit stronger connections. Specifically, there was a significant overlap in attributes between Jianggan men (point L) and Jianggan women (point K), indicating a close relationship. In addition, a strong association exists between Jianggan women (point K) and Wuyi women (point I).

We summarized some information after a thorough semantic analysis. All age groups exhibited low performance in actions,

including using fluoride toothpaste, dental floss, and scheduling timely visits to an oral hospital. Thus, these actions should be appreciated and strengthened. Middle-aged individuals and older adults still have poor health knowledge and awareness of health care. It is necessary to disseminate and strengthen certain cognitions, such as the knowledge that fossa closure and fluoride can protect teeth.

User Study

To further evaluate the effectiveness of our system, we invited 20 undergraduate and graduate students (12 male students and 8 female students) in digital media technology to participate in a user study. We first introduce the purpose and features of this system and then teach students how to use it. Typically, users need only 10 to 15 minutes of training time to understand the meaning of each view and the function of our system. Afterward, they were asked to perform a series of tasks over a defined period, which were closely related to the analysis tasks in *Methods* section. The specific tasks were as follows:

- Disease
  - Task 1.1. Which disease is more prevalent in Jianggan than in other areas?
  - Task 1.2. Which gender has the highest prevalence of AL12+ disease?
- Group
  - Task 2.1. What are the 3 main behavioral characteristics of the male population aged 55 to 64 years in Jianggan?
  - Task 2.2. Which area has a higher prevalence of DRoot among girls aged 12 to 15 years?
- Correlation
  - Task 3.1. Is GBI positively or negatively correlated with use of the toothpick (UT)?
  - Task 3.2. Which disease is most likely to present with symptoms of difficulty chewing?
- Semantics
  - Task 4.1. What are the groups with severe DT (DT4)?
  - Task 4.2. What are the common characteristics of the women aged 55 to 64 years in Luqiao and Wenling?

To further demonstrate the effectiveness of the system, users will perform the task twice, once without the system and once with it. When the system was not applicable, we provided the users with a condensed version of the surveys. For each experiment, we set the maximum completion time to 90 seconds. Table 1 and Figure 6 record the percentage of the users able to complete the task correctly in a given time and the average and SD of the time completed.

Figure 6. User study results. (A) Comparison of completion time. (B) Completion of accuracy rate. DT: decayed tooth system for data mining; OUR: use our system for data mining; SUR: data mining without our system.

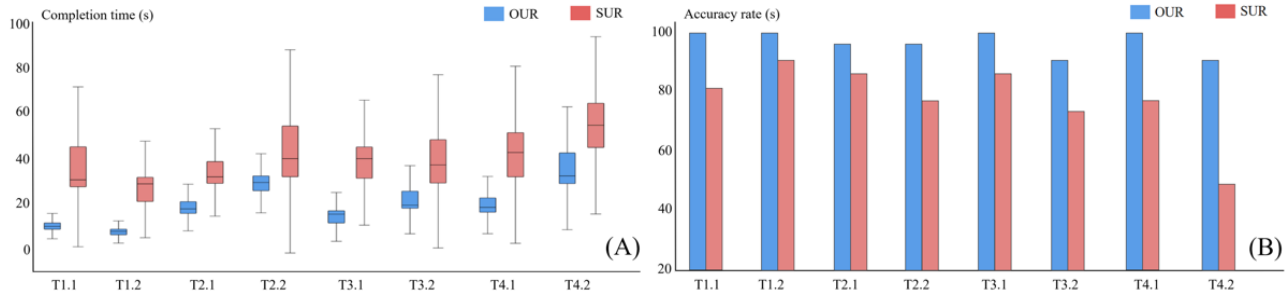


Table 1 presents completion time, average completion time, and standard completion time for both cases. Figure 6 displays specific completion time and accuracy rate through a box chart and a bar chart. It is obvious that using this system would result in more accurate and efficient exploration of oral health surveys. In addition, we collected user feedback after performing the above tasks. They all agreed that the system was quite intriguing and may give them a more intuitive understanding of oral health.

Table 1. User study results.

Category and user task	Accuracy rate (%)		Average completion times (s)		Standard completion times (s)	
	OUR[a]	SUR[b]	OUR	SUR	OUR	SUR
Disease						
T1.1	100	80	10.47	38.33	10.1	30.45
T1.2	100	90	8.2	26.99	8.1	28.68
Group						
T2.1	95	85	19.05	33.55	17.75	31.8
T2.2	95	75	29.35	42.29	29.25	39.7
Correlation						
T3.1	100	85	14.54	38.9	15.3	39.2
T3.2	90	70	20.85	39.21	19.25	36.75
Semantics						
T4.1	100	75	19.06	42.29	18.15	42
T4.2	90	45	34.73	54.48	31.8	53.8

[a]OUR: use our system for data mining.

[b]SUR: data mining without our system.

Expert Interview

After domain experts used this system to examine oral survey data, we conducted a semistructured interview to collect their opinions on system capability, visual design, and interaction.

System Capability and Effectiveness

The experts expressed their appreciation for the functions provided by this system. They concluded that the system makes it possible for both experts and regular users to quickly and intuitively perceive the initially complex and laborious large-scale oral survey data as well as to easily compare the characteristics of various diseases and groups. Moreover, they agreed that the system effectively revealed the correlation among diseases, behaviors, symptoms, and cognitions. In particular, the oral survey data were transformed into a knowledge graph, a novel approach that is not commonly used in daily survey data analysis. By leveraging the knowledge graph and its query function, this breakthrough enables researchers to go beyond traditional methods that focus on specific tasks and features. It allows them to comprehend large-scale data with complex semantic patterns, making it easier to understand. Ultimately, it greatly enhances their insights into the data.

Visual Design and Interactions

Domain experts praised the user-friendly interface and the well-designed system, aligning each view with its respective function and interaction. The layout facilitates rich interactions, correlation discovery, semantic analysis, and attribute feature exploration. Users can easily comprehend and use them without prior knowledge. One of the experts recognized the scatterplot matrix view's value in conveniently analyzing group comparison and attribute correlation simultaneously. Another expert emphasized the search capability of the knowledge graph, which allowed him to independently examine valuable information that was difficult to find in regular visual graphs. He suggested that it would be better if the system could offer label options or prompts to explore the nodes. Overall, the experts evaluated the system's integration of visualizations and interactions, offering a comprehensive range of intelligent explorations for oral health surveys.

Discussion

Principal Findings

A series of studies and experiments have demonstrated that our system can help users understand their oral health conditions and conduct in-depth data exploration and analysis. Furthermore, we conducted a thorough investigation of the visualization analysis tools to compare them with our system. We found that existing tools and libraries provide a rich set of plotting capabilities. However, the visualization analysis tools used in our study are primarily oriented toward specific tasks to visually present and obtain deep insights into oral health surveys. It is implemented using web-based technologies such as the D3.js visualization framework, which offers greater flexibility and customization options for analyzing oral health report data. Existing tools related to oral health analysis mostly include 3D digital dental model software and oral x-ray image processing software, which provide detailed visualization of dental structures. Nevertheless, these tools fail to capture the broader context of oral health such as group characteristics and disease patterns. Moreover, they often provide simple chart-based visualizations, such as pie charts and bar charts, lacking the personalized visualization design and interactive features essential for the comprehensive examination of intricate data. Therefore, it was concluded that our system allows for more customized visualizations based on specific requirements, facilitating a more detailed analysis of oral health surveys.

Overall, our system has specific advantages compared with other analysis tools; however, there are also some issues that

are not well solved, which will be addressed in future work. (1) Scalability is the major concern of this system. The current design in the scatterplot matrix view displays up to 8 attributes and 6 groups simultaneously, whereas the struct view shows up to 7 diseases simultaneously. Even if we set the interaction or scrolling function in it, it still imposes a heavy memory burden on users. Therefore, in the future, we intend to tackle the problem of how to show information more effectively in a limited screen space. (2) Despite its ability to analyze various factors, such as groups, diseases, behaviors, and other attributes in existing data, it currently lacks the capability to predict oral health for new groups or individuals. Combining the deep learning model and oral health professional knowledge, learning from existing multidimensional surveys, and predicting the prevalence of unknown groups will be the focus of future work. (3) In this study, the oral health sample was limited to Zhejiang Province, with a small-scale and narrow regional span, resulting in insufficient group differences. Future studies should consider the multidimensional feature of the disease to explore more robust results. We plan to expand our data collection by conducting oral surveys in additional regions, enabling a more comprehensive exploration of oral health from various dimensions.

Conclusions

In this study, we proposed a visualization framework for multidimensional oral health surveys. We drew data portraits for diseases and groups based on multidimensional attributes. Then, we built correlation patterns for diseases, behaviors, symptoms, and cognitions to reveal their correlation features. On the basis of the extricated knowledge of diseases, groups, behaviors, and other attributes, a knowledge graph is provided to reveal the semantic information. A series of user-friendly interactions are integrated to propose a visual analysis system that can help users further explore the regulations of oral health conditions. Case studies based on real-world data sets demonstrate the effectiveness of our system in the exploration of oral diseases, thereby offering enhanced data analysis capabilities and decision support for health care teams.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (62277013 and 62177040), National Statistical Science Research Project (2022LY099), Zhejiang Provincial Natural Science Foundation of China (LTGG23H260003), the China Oral Health Foundation (A2021-008), and Zhejiang Statistical Science Research Project.

Conflicts of Interest

None declared.

References

1. Fisher J, Selikowitz HS, Mathur M, Varenne B. Strengthening oral health for universal health coverage. Lancet 2018 Sep;392(10151):899-901. [doi: 10.1016/s0140-6736(18)31707-0]
2. Pattanaik S, John MT, Kohli N, Davison ML, Chung S, Self K, et al. Item and scale properties of the Oral Health Literacy Adults Questionnaire assessed by item response theory. J Public Health Dent 2021 Sep 10;81(3):214-223. [doi: 10.1111/jphd.12434] [Medline: 33305385]
3. Heisel MJ, Flett GL. The Social Hopelessness Questionnaire (SHQ): psychometric properties, distress, and suicide ideation in a heterogeneous sample of older adults. J Affect Disord 2022 Feb 15;299:475-482. [doi: 10.1016/j.jad.2021.11.021] [Medline: 34774647]
4. Drapała J, Żatuchin D, Sobecki J. Multidimensional data visualization applied for user's questionnaire data quality assessment. In: Agent and Multi-Agent Systems: Technologies and Applications. Berlin, Heidelberg: Springer; 2010.
5. Zhang X, Klawonn F, Grigull L, Lechner W. VoQs: a web application for visualization of questionnaire surveys. In: Advances in Intelligent Data Analysis XIV. Cham: Springer; 2015.
6. Atkinson NL, Saperstein SL, Pleis J. Using the internet for health-related activities: findings from a national probability sample. J Med Internet Res 2009 Feb 20;11(1):e4 [FREE Full text] [doi: 10.2196/jmir.1035] [Medline: 19275980]
7. Powell J, Inglis N, Ronnie J, Large S. The characteristics and motivations of online health information seekers: cross-sectional survey and qualitative interview study. J Med Internet Res 2011 Feb 23;13(1):e20 [FREE Full text] [doi: 10.2196/jmir.1600] [Medline: 21345783]
8. O'Brien KK, Solomon P, Worthington C, Ibáñez-Carrasco F, Baxter L, Nixon SA, HIV, Health And Rehabilitation Survey Catalyst Team. Considerations for conducting web-based survey research with people living with human immunodeficiency virus using a community-based participatory approach. J Med Internet Res 2014 Mar 13;16(3):e81 [FREE Full text] [doi: 10.2196/jmir.3064] [Medline: 24642066]
9. Aggarwal R, Farag S, Martin G, Ashrafian H, Darzi A. Patient perceptions on data sharing and applying artificial intelligence to health care data: cross-sectional survey. J Med Internet Res 2021 Aug 26;23(8):e26162 [FREE Full text] [doi: 10.2196/26162] [Medline: 34236994]
10. Nakamura C, Bromberg M, Bhargava S, Wicks P, Zeng-Treitler Q. Mining online social network data for biomedical research: a comparison of clinicians' and patients' perceptions about amyotrophic lateral sclerosis treatments. J Med Internet Res 2012 Jun 21;14(3):e90 [FREE Full text] [doi: 10.2196/jmir.2127] [Medline: 22721865]

11. Zhou Z, Zhang X, Guo Z, Liu Y. Visual abstraction and exploration of large-scale geographical social media data. Neurocomputing 2020 Feb;376:244-255. [doi: 10.1016/j.neucom.2019.10.072]
12. Scheepens R, van de Wetering H, van Wijk JJ. Non-overlapping aggregated multivariate glyphs for moving objects. In: Proceedings of the IEEE Pacific Visualization Symposium. 2014 Presented at: IEEE Pacific Visualization Symposium; Mar 04-07, 2014; Yokohama, Japan. [doi: 10.1109/PacificVis.2014.13]
13. Brehmer M, Lee B, Isenberg P, Choe EK. A comparative evaluation of animation and small multiples for trend visualization on mobile phones. IEEE Trans Visual Comput Graphics 2020 Jan;26(1):364-374. [doi: 10.1109/tvcg.2019.2934397]
14. Viau C, McGuffin MJ, Chiricota Y, Jurisica I. The FlowVizMenu and parallel scatterplot matrix: hybrid multidimensional visualizations for network exploration. IEEE Trans Visual Comput Graphics 2010 Nov;16(6):1100-1108. [doi: 10.1109/tvcg.2010.205]
15. Chen H, Engle S, Joshi A, Ragan ED, Yuksel BF, Harrison L. Using animation to alleviate overdraw in multiclass scatterplot matrices. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. 2018 Presented at: CHI '18: CHI Conference on Human Factors in Computing Systems; Apr 21-26, 2018; Montreal, QC, Canada. [doi: 10.1145/3173574.3173991]
16. Zhou Z, Ye Z, Yu J, Chen W. Cluster-aware arrangement of the parallel coordinate plots. J Visual Languages Comput 2018 Jun;46:43-52. [doi: 10.1016/j.jvlc.2017.10.003]
17. Zhou Z, Ma Y, Zhang Y, Liu Y, Liu Y, Zhang L, et al. Context-aware visual abstraction of crowded parallel coordinates. Neurocomputing 2021 Oct;459:23-34. [doi: 10.1016/j.neucom.2021.05.005]
18. Rao R, Card SK. The table lens: merging graphical and symbolic representations in an interactive focus+context visualization for tabular information. In: Proceedings of the Conference Companion on Human Factors in Computing Systems. 1994 Presented at: CHI94: ACM Conference on Human Factors in Computer Systems; Apr 24-28, 1994; Boston, Massachusetts, USA. [doi: 10.1145/259963.260391]
19. Keim DA, Kriegel HP. VisDB: database exploration using multidimensional visualization. IEEE Comput Grap Appl 1994 Sep;14(5):40-49. [doi: 10.1109/38.310723]
20. Xia J, Chen T, Zhang L, Chen W, Chen Y, Zhang X, et al. SMAP: a joint dimensionality reduction scheme for secure multi-party visualization. In: Proceedings of the IEEE Conference on Visual Analytics Science and Technology (VAST). 2020 Presented at: IEEE Conference on Visual Analytics Science and Technology (VAST); Oct 25-30, 2020; Salt Lake City, UT, USA. [doi: 10.1109/vast50239.2020.00015]
21. Xia J, Huang L, Lin W, Zhao X, Wu J, Chen Y, et al. Interactive visual cluster analysis by contrastive dimensionality reduction. IEEE Trans Visual Comput Graphics 2022;29(1):734-744. [doi: 10.1109/tvcg.2022.3209423]
22. Donath J, Dragulescu A, Zinman A, Viégas F, Xiong R. Data portraits. Leonardo 2010 Aug;43(4):375-383. [doi: 10.1162/leon_a_00011]
23. Xiong R, Donath J. PeopleGarden: creating data portraits for users. In: Proceedings of the 12th annual ACM symposium on User interface software and technology. 1999 Presented at: UIST99: Twelfth Annual Symposium on User Interface Software and Technology; Nov 7-10, 1999; Asheville, North Carolina, USA. [doi: 10.1145/320719.322581]
24. He P, Niu H, Sun Z, Li T. Accounting index of COVID-19 impact on Chinese industries: a case study using big data portrait analysis. Emerg Market Finance Trade 2020 Jul 25;56(10):2332-2349. [doi: 10.1080/1540496x.2020.1785866]
25. Ji S, Pan S, Cambria E, Marttinen P, Yu PS. A survey on knowledge graphs: representation, acquisition, and applications. IEEE Trans Neural Netw Learn Syst 2022 Feb;33(2):494-514. [doi: 10.1109/tnnls.2021.3070843]
26. Zheng S, Wang F, Bao H, Hao Y, Zhou P, Xu B. Joint extraction of entities and relations based on a novel tagging scheme. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017 Presented at: 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Jul 30-Aug 4, 2017; Vancouver, Canada. [doi: 10.18653/v1/P17-1113]
27. Liu F, Chen M, Roth D, Collier N. Visual pivoting for (unsupervised) entity alignment. Proc AAAI Conf Artif Intel 2021 May 18;35(5):4257-4266. [doi: 10.1609/aaai.v35i5.16550]
28. Wang X, Wang D, Xu C, He X, Cao Y, Chua T. Explainable reasoning over knowledge graphs for recommendation. Proc AAAI Conf Artif Intel 2019 Jul 17;33(01):5329-5336. [doi: 10.1609/aaai.v33i01.33015329]
29. Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C. Neural architectures for named entity recognition. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016 Presented at: 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; Jun 12-17, 2016; San Diego, California. [doi: 10.18653/v1/n16-1030]
30. Chen Y, Goldberg S, Wang DZ, Johri SS. Ontological pathfinding: mining first-order knowledge from large knowledge bases. In: Proceedings of the 2016 International Conference on Management of Data. 2016 Presented at: SIGMOD/PODS'16: International Conference on Management of Data; Jun 26-Jul 1, 2016; San Francisco, California, USA.
31. Wang CC, Cheng PJ. Translating representations of knowledge graphs with neighbors. In: Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. 2018 Presented at: SIGIR '18: The 41st International ACM SIGIR conference on research and development in Information Retrieval; Jul 8-12, 2018; Ann Arbor, MI, USA. [doi: 10.1145/3209978.3210085]

32. Vashishth S, Sanyal S, Nitin V, Agrawal N, Talukdar P. InteractE: improving convolution-based knowledge graph embeddings by increasing feature interactions. Proc AAAI Conf Artif Intel 2020 Apr 03;34(03):3009-3016. [doi: 10.1609/aaai.v34i03.5694]
33. Wang X, Xu Y, He X, Cao Y, Wang M, Chua TS. Reinforced negative sampling over knowledge graph for recommendation. In: Proceedings of The Web Conference 2020. 2020 Presented at: WWW '20: The Web Conference 2020; Apr 20-24, 2020; Taipei, Taiwan. [doi: 10.1145/3366423.3380098]
34. Zhang Y, Dai H, Kozareva Z, Smola A, Song L. Variational reasoning for question answering with knowledge graph. Proc AAAI Conf Artif Intel 2018 Apr 26;32(1). [doi: 10.1609/aaai.v32i1.12057]
35. Xiong C, Power R, Callan J. Explicit semantic ranking for academic search via knowledge graph embedding. In: Proceedings of the 26th International Conference on World Wide Web. 2017 Presented at: WWW '17: 26th International World Wide Web Conference; Apr 3-7, 2017; Perth, Australia. [doi: 10.1145/3038912.3052558]
36. Li H, Wang Y, Zhang S, Song Y, Qu H. KG4Vis: a knowledge graph-based approach for visualization recommendation. IEEE Trans Vis Comput Graph 2022 Jan;28(1):195-205. [doi: 10.1109/TVCG.2021.3114863] [Medline: 34587080]
37. Sousa D, Couto FM. Biomedical relation extraction with knowledge graph-based recommendations. IEEE J Biomed Health Inform 2022 Aug;26(8):4207-4217. [doi: 10.1109/JBHI.2022.3173558] [Medline: 35536818]
38. Tang Y, Han H, Yu X, Zhao J, Liu G, Wei L. An intelligent question answering system based on power knowledge graph. In: Proceedings of the IEEE Power & Energy Society General Meeting (PESGM). 2021 Presented at: IEEE Power & Energy Society General Meeting (PESGM); Jul 26-29, 2021; Washington, DC, USA. [doi: 10.1109/pesgm46819.2021.9638018]
39. Latif S, Agarwal S, Gottschalk S, Chrosch C, Feit F, Jahn J, et al. Visually connecting historical figures through event knowledge graphs. In: Proceedings of the IEEE Visualization Conference (VIS). 2021 Presented at: IEEE Visualization Conference (VIS); Oct 24-29, 2021; New Orleans, LA, USA. [doi: 10.1109/vis49827.2021.9623313]
40. Wang H, Chen H, Zhu H, Zhou N. Oral Health Status Survey and Prevention of Common Diseases in Zhejiang Province. Zhejiang, China: Zhejiang University Press; 2019.

Abbreviations

AL: attachment loss
DRoot: decayed teeth root due to caries
DT: decayed tooth
ft: fill deciduous tooth due to caries
FT: fill tooth due to caries
GBI: bleeding gums index
MT: missing tooth
PDD: periodontal pocket depth

Edited by J Hefner; submitted 05.02.23; peer-reviewed by MC Skelton, M Pang; comments to author 10.04.23; revised version received 28.05.23; accepted 31.05.23; published 01.08.23.

Please cite as:
Xu T, Ma Y, Pan T, Chen Y, Liu Y, Zhu F, Zhou Z, Chen Q
Visual Analytics of Multidimensional Oral Health Surveys: Data Mining Study
JMIR Med Inform 2023;11:e46275
URL: https://medinform.jmir.org/2023/1/e46275
doi:10.2196/46275
PMID:37526971

Original Paper

Analyzing and Forecasting Pediatric Fever Clinic Visits in High Frequency Using Ensemble Time-Series Methods After the COVID-19 Pandemic in Hangzhou, China: Retrospective Study

Wang Zhang[1,2,3*], MS; Zhu Zhu[1,2,3*], PhD; Yonggen Zhao[1,2,3], BE; Zheming Li[1,2,3], BE; Lingdong Chen[1,2,3], ME; Jian Huang[1,2,3], MS; Jing Li[1,2,3], MS; Gang Yu[1,2,3], ME

[1]Department of Data and Information, Children's Hospital, Zhejiang University School of Medicine, Hangzhou, China

[2]Sino-Finland Joint AI Laboratory for Child Health of Zhejiang Province, Hangzhou, China

[3]National Clinical Research Center for Child Health, Hangzhou, China

[*]these authors contributed equally

Corresponding Author:
Gang Yu, ME
Department of Data and Information
Children's Hospital, Zhejiang University School of Medicine
No. 3333 Binsheng Road, Binjiang District
Hangzhou, 310000
China
Phone: 86 13588773370
Email: yugbme@zju.edu.cn

Abstract

Background: The COVID-19 pandemic has significantly altered the global health and medical landscape. In response to the outbreak, Chinese hospitals have established 24-hour fever clinics to serve patients with COVID-19. The emergence of these clinics and the impact of successive epidemics have led to a surge in visits, placing pressure on hospital resource allocation and scheduling. Therefore, accurate prediction of outpatient visits is essential for informed decision-making in hospital management.

Objective: Hourly visits to fever clinics can be characterized as a long-sequence time series in high frequency, which also exhibits distinct patterns due to the particularity of pediatric treatment behavior in an epidemic context. This study aimed to build models to forecast fever clinic visit with outstanding prediction accuracy and robust generalization in forecast horizons. In addition, this study hopes to provide a research paradigm for time-series forecasting problems, which involves an exploratory analysis revealing data patterns before model development.

Methods: An exploratory analysis, including graphical analysis, autocorrelation analysis, and seasonal-trend decomposition, was conducted to reveal the seasonality and structural patterns of the retrospective fever clinic visit data. The data were found to exhibit multiseasonality and nonlinearity. On the basis of these results, an ensemble of time-series analysis methods, including individual models and their combinations, was validated on the data set. Root mean square error and mean absolute error were used as accuracy metrics, with the cross-validation of rolling forecasting origin conducted across different forecast horizons.

Results: Hybrid models generally outperformed individual models across most forecast horizons. A novel model combination, the hybrid neural network autoregressive (NNAR)-seasonal and trend decomposition using Loess forecasting (STLF), was identified as the optimal model for our forecasting task, with the best performance in all accuracy metrics (root mean square error=20.1, mean absolute error=14.3) for the 15-days-ahead forecasts and an overall advantage for forecast horizons that were 1 to 30 days ahead.

Conclusions: Although forecast accuracy tends to decline with an increasing forecast horizon, the hybrid NNAR-STLF model is applicable for short-, medium-, and long-term forecasts owing to its ability to fit multiseasonality (captured by the STLF component) and nonlinearity (captured by the NNAR component). The model identified in this study is also applicable to hospitals in other regions with similar epidemic outpatient configurations or forecasting tasks whose data conform to long-sequence time series in high frequency exhibiting multiseasonal and nonlinear patterns. However, as external variables and disruptive events were not accounted for, the model performance declined slightly following changes in the COVID-19 containment policy in China. Future work may seek to improve accuracy by incorporating external variables that characterize moving events or other factors as well as by adding data from different organizations to enhance algorithm generalization.

(JMIR Med Inform 2023;11:e45846) doi:10.2196/45846

KEYWORDS

time-series forecasting; outpatient visits; hospital management; pediatric fever clinic; long sequence; visits in high frequency; COVID-19

Introduction

Background

COVID-19 is the most severe global pandemic of the 21st century, which has brought major changes to the global health care environment [1]. According to statistics from the World Health Organization, there have been >760 million confirmed cases of COVID-19 worldwide, including nearly 7 million deaths to date. Although the World Health Organization has declared that COVID-19 no longer constitutes a "public health emergency of international concern," it remains a serious infectious disease that will persist for the foreseeable future. Moreover, since the onset of the epidemic, numerous epidemic infections have also emerged, including influenza A, respiratory syncytial virus infection, and mycoplasma pneumonia. Successive waves of respiratory infections led to a significant increase in the number of patients presenting with fever. This prompted governments and hospitals to take measures for patient management to prevent viral transmission and control the risk of hospital-acquired infections [2,3].

In China, since the outbreak of COVID-19, most public hospitals have established fever clinics that can achieve individual closed-loop management in the clinics themselves (as shown in Figure 1). This enables the centralized treatment of patients with fever infections. As mandated by the National Health Commission, the allocation of resources and the operation in fever clinics must strictly adhere to established guidelines [4] to minimize the risk of hospital infection. Some hospitals even operate their fever clinics 24/7. Despite the relaxation of China's epidemic containment policy since the end of 2022, many hospitals continue to operate their fever clinics as usual.

The presence of epidemics and the establishment of continuously operating fever clinics are altering visitation patterns, particularly among pediatric patients. Compared with adult patients, pediatric patients require more attentive care from both their guardians and medical staff, and their conditions are more prone to relapse, resulting in a continuous and intensive trend of fever clinic visits among pediatric patients during the pandemic. This poses great challenges for hospital outpatient management and the prevention of hospital-acquired infections. In this context, there is growing interest in the study of outpatient visit forecasting.

Forecasting visits to fever clinics offers numerous benefits for hospital management. Accurate and timely visit forecasts can facilitate the rational allocation of manpower and medical consumables in outpatient departments as well as the refined management and scheduling of medical equipment and facilities. Moreover, hour-level outpatient visit forecasts can provide a valuable decision-making reference for patients' time management, enhancing efficiency for both hospitals and patients. Hourly visits to fever clinics can be characterized as a long-sequence time series in high frequency owing to their high sampling rate and large time window. The unique visitation patterns of children with fever will also certainly be reflected in the time-series data. Therefore, a peer-to-peer approach capable of uncovering the intrinsic patterns of pediatric fever clinic visit time series and establishing accurate and fine-grained visit forecast models is highly desirable.

Figure 1. Fever clinic deployment instructions in the designated hospital.

Related Work

Time-series analysis and forecasting have been widely applied in various fields, such as disease analysis [5-9], hospital operation management [10-14], and drug management [15-18]. Numerous studies have been conducted to forecast daily and hourly arrivals or occupancies in emergency departments. Hertzum [19] and Choudhury and Urena [20] effectively predicted hourly arrivals using autoregressive integrated moving average (ARIMA) models, whereas Becerra et al [21] studied a seasonal ARIMA (SARIMA) forecast model based on daily emergency admissions for respiratory outpatients. Cheng et al [22] and Whitt and Zhang [23] applied SARIMA with an external repressor to forecast hourly occupancy. Deep learning algorithms, such as the variational autoencoder proposed by Harrou et al [24] and the long short-term memory (LSTM) used by Etu et al [25], have also been applied to such problems, demonstrating that prediction results can significantly aid decision support in hospital management. Zhang et al [26] and Sudarshan et al [27] incorporated external variables, such as calendar and meteorological information, into the LSTM model, verifying the improvement in the accuracy of the models established in their research case.

Khaldi et al [28] investigated a hybrid model combining artificial neural networks with ensemble empirical mode decomposition, which exhibited better approximation and generalization capabilities than the benchmarking models when applied to weekly arrivals in emergency departments. Deng et al [29] proposed a hybrid ARIMA-LSTM model optimized by the backpropagation neural networks that achieved more accurate and stable predictions than the respective single models and the traditional hybrid model when forecasting weekly and monthly outpatient visits to the respiratory department. Perone [30] used single-step time-series methods and their feasible ensembles to forecast the arrival of hospitalized patients with COVID-19 presenting with mild symptoms, as well as those in intensive care units. They discovered that hybrid models were significantly better at capturing linear, nonlinear, and seasonal pandemic patterns than their respective single models on both time series.

Although numerous studies have been dedicated to forecasting outpatient visits, the time-series models used in existing hospital visit forecasting studies are limited in their ability, for they handle only a single seasonality pattern. When applied to long-sequence time series in high frequency, these models are unable to capture all seasonal patterns present in the data, resulting in the loss of data features and increased difficulty in forecasting. This is the significant drawback of single-seasonal-pattern models. Additionally, although cross-validation has been conducted in previous time-series studies to evaluate model performance, it has rarely been used as a basis for comparing the forecasting performance of models across different forecast horizons, leading to an incomplete analysis and evaluation of the predictive capabilities of different models. This study aimed to address these limitations.

Objective

In this study, we aimed to develop a reliable forecasting model for the hospital management of fever clinics. Unlike previous studies, we focused on establishing models suitable for long-sequence time series in high frequency. The model needed to meet the following 3 requirements to facilitate our fever clinic visit forecast task. Firstly, we prioritized hourly forecasts over daily forecasts to predict fever clinic visits, as this allows for greater flexibility in hospital management and resource allocation. Second, an ideal model should be capable of making medium- to long-term forecasts (eg, 15 days) on high-frequency time-series data to aid hospital managers in making informed

decisions. Finally, the model should be scalable and stable and exhibit robust performance.

To this end, we first conducted an exploratory analysis of time-series data to uncover the patterns and features of fever clinic hourly visit time series and then used an ensemble of time-series models and their combinations for fever clinic visit forecasting. For model evaluation, we used cross-validation to compare the accuracy of all models across different forecast horizons and analyzed the results. Exploratory analysis can help discover the inherent laws of data, which makes it easier to find models that fit the characteristics of the data. Cross-validation across different forecast horizons helps find models with superior scalability and stability. These are the 2 points that represent the main highlights of our research compared with existing work.

Methods

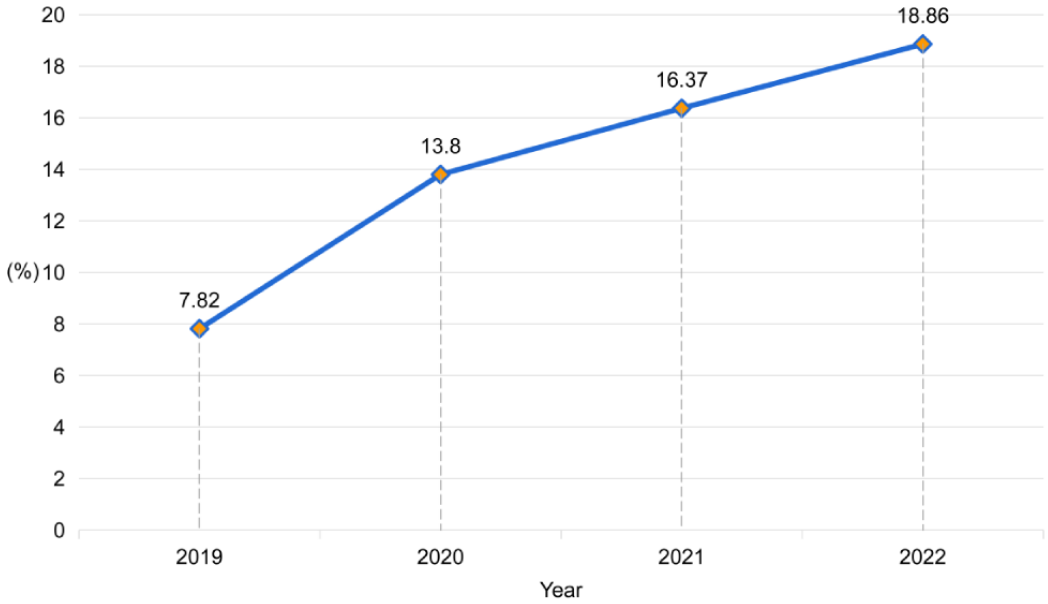
Ethical Considerations

The authors are accountable for all aspects of the work and are responsible for ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. This study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The study was approved by the Academic Ethics Committee of the Children's Hospital of Zhejiang University School of Medicine (2020-RIB-058). The requirement for individual consent for this retrospective analysis was waived.

Study Participants

This study focused on the Children's Hospital of Zhejiang University School of Medicine, a preeminent comprehensive class A tertiary children's hospital located in the Zhejiang province, China. In response to the COVID-19 outbreak in 2020, the hospital's fever clinic transitioned to a 24-hour emergency operation mode, providing uninterrupted care for pediatric patients. The fever clinic operates as an autonomous department fully equipped with comprehensive medical resources. Figure 2 presents data illustrating the proportion of fever clinic visits relative to the total number of outpatient visits over the past 4 years. The data revealed a consistent increase in the proportion of fever clinic visits, with a notable surge in 2020.

Figure 2. The ratio of fever clinic visits to the total number of outpatient visits over the past 4 years.



Data Collection and Preprocessing

The data set used in this study consisted of authentic data extracted from the electronic medical record (EMR) text of the aforementioned hospital's fever clinic. Data were collected from January 23, 2020, to May 23, 2023, encompassing the onset of the COVID-19 outbreak and the phases of strict containment policy (January 2021 to November 2022) and open containment policy (December 2022 to present) implemented in the region. We tallied hourly fever clinic visits based on patients' initial visit records in the EMR and divided each day's 24-hour data into 24 nonoverlapping segments. Our data set comprises a time series of 29,208 data points (1217 days × 24 hours) representing hourly visits. Our data set exhibited a paucity of missing data, which we addressed by filling sporadic gaps in the hourly count with a value of 0. Outliers were identified through the remainder sequence obtained via seasonal-trend decomposition using Loess (STL) decomposition and defined as values exceeding 3 IQRs from the central 50% of the data. These outliers were subsequently smoothed using linear interpolation.

To further scrutinize our data set, we conducted statistical analyses based on children's developmental stages as determined by their educational level [31]. The total number of patients was 1,590,909, and we divided them into the following 5 groups: infants (aged 0-2 years; n=661,425, 41.58%), preprimary children (aged 3-5 years; n=599,464, 37.68%), primary school children (aged 6-11 years; n=302,046, 18.99%), junior secondary students (aged 12-14 years; n=25,364, 1.59%), and senior secondary students (aged 15-17 years; n=2610, 0.16%). Table 1 depicts fever clinic visits for each group, revealing that the patients were predominantly infants and preprimary children.

Furthermore, we extracted International Classification of Diseases, 10th Revision disease diagnosis codes [32] from the EMR data and compiled statistics on related diseases in the fever clinic. Our results indicate that respiratory diseases constitute the largest proportion (738,635/1,590,909, 46.43%) of fever clinic cases, followed by infectious and parasitic diseases (347,199/1,590,909, 21.82%) and digestive diseases (174,814/1,590,909, 10.99%), all of which are influenced by climate change.

Table 1. Distribution of patients who visited the fever clinic across age groups from January 23, 2020, to May 23, 2023.

	0-2 years	3-5 years	6-11 years	12-14 years	15-17 years
Girls, n	297,590	283,567	141,768	10,887	1221
Boys, n	363,835	315,897	160,278	14,477	1389

Exploratory Data Analysis

Before constructing forecasting models, it is imperative to comprehend the behavior of data in the time domain. Using statistical graphics and data visualization techniques, time-series patterns can be extracted and interpreted, facilitating model selection and minimizing errors.

Graphic Analysis

Figure 3 depicts the hourly fever clinic visit time series from our data set. We plotted time-series data before and after the changes in the epidemic containment policy, with December 19, 2022, as the separation point. To identify the underlying patterns in the time series, we analyzed the data from a seasonal perspective. Given that our data are hourly, they may exhibit 3 types of seasonality: daily, weekly, and yearly. These patterns are plotted in Figures 4-6.

As illustrated in Figure 4, the diurnal pattern of fever clinic visits exhibited relatively fewer visits during the early morning hours, with 3 prominent peaks occurring at 9 AM, 2 PM, and 8 PM, indicating heightened visitation. Figure 5 reveals that visitation peaks are most pronounced on Mondays and Tuesdays, diminishing on Wednesdays and Thursdays before a resurgence from Friday through the weekend. Despite an overall decline in fever clinic visits in 2023 due to the relaxation of COVID-19 policies, the time series which consists of data spanning over 3 years still exhibits clear annual periodicity. As shown in Figure 6, values fluctuate systematically with seasonal and even monthly variations, with elevated values during winter, diminished values during spring, and peak outpatient periods coinciding with summer vacation. The location of troughs in Figure 6 appears to be influenced by movable holidays such as the Chinese New Year. From Figures 4-6, we can deduce that the time series exhibits conspicuous multiseasonal patterns, including daily, weekly, and yearly seasonalities, thereby exhibiting robust predictability.

Figure 3. Time-series plot of hourly fever clinic visits from January 23, 2020, to May 23, 2023.

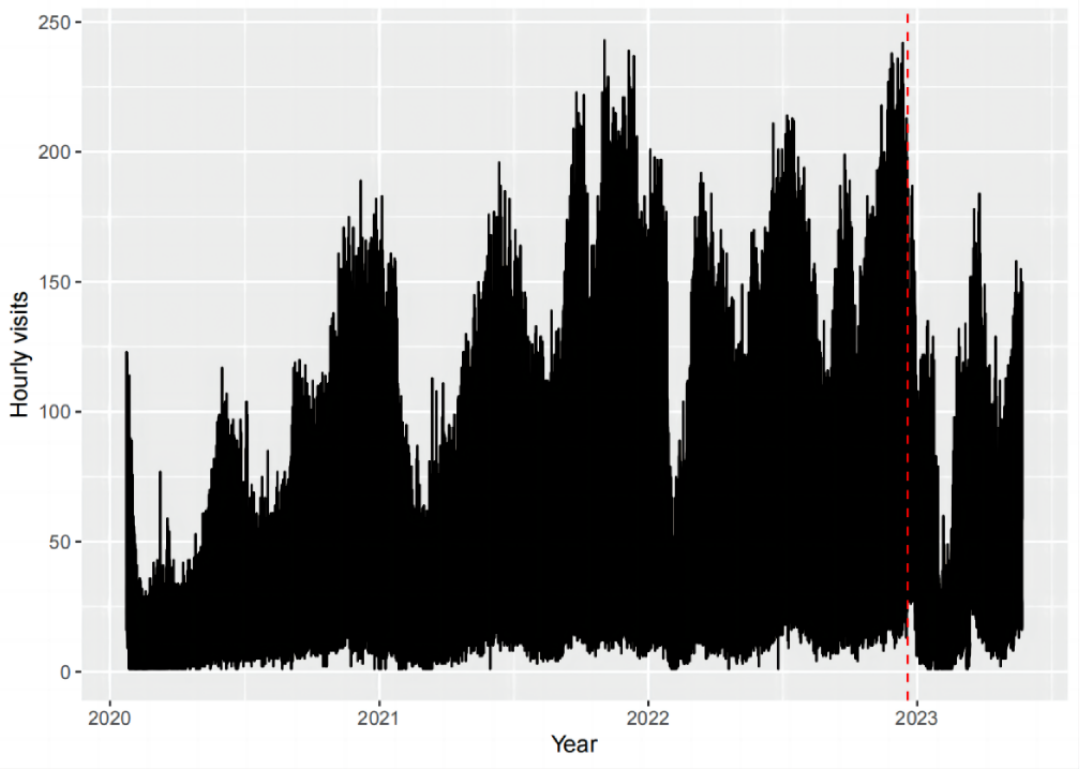


Figure 4. Daily seasonal patterns in the hourly fever clinic visit time series.

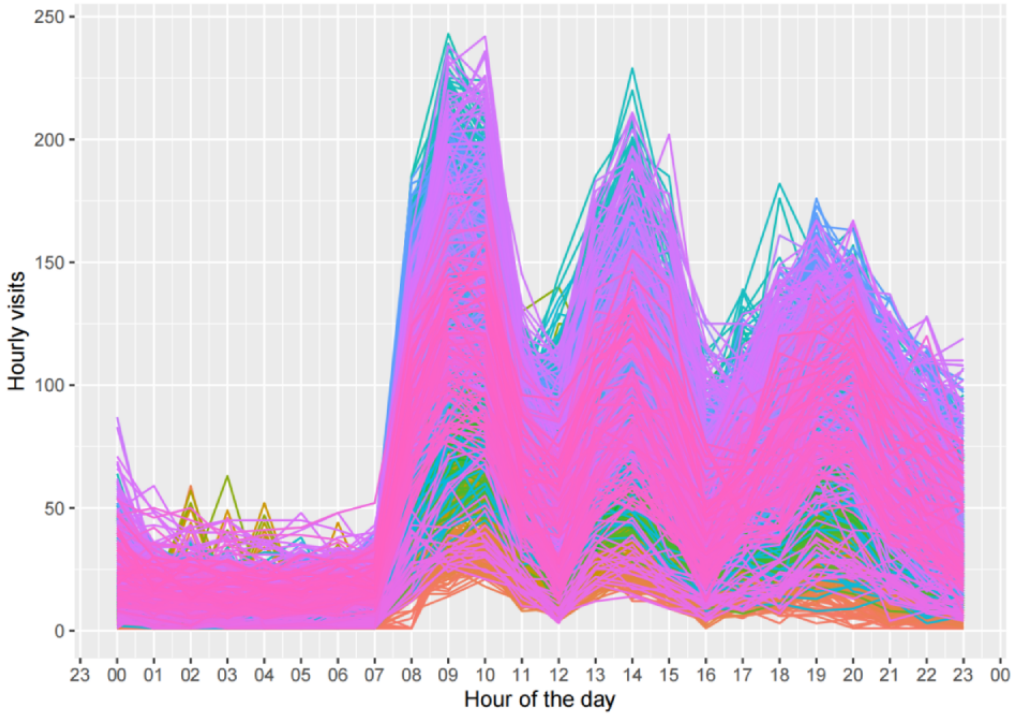


Figure 5. Weekly seasonal patterns in the hourly fever clinic visit time series.

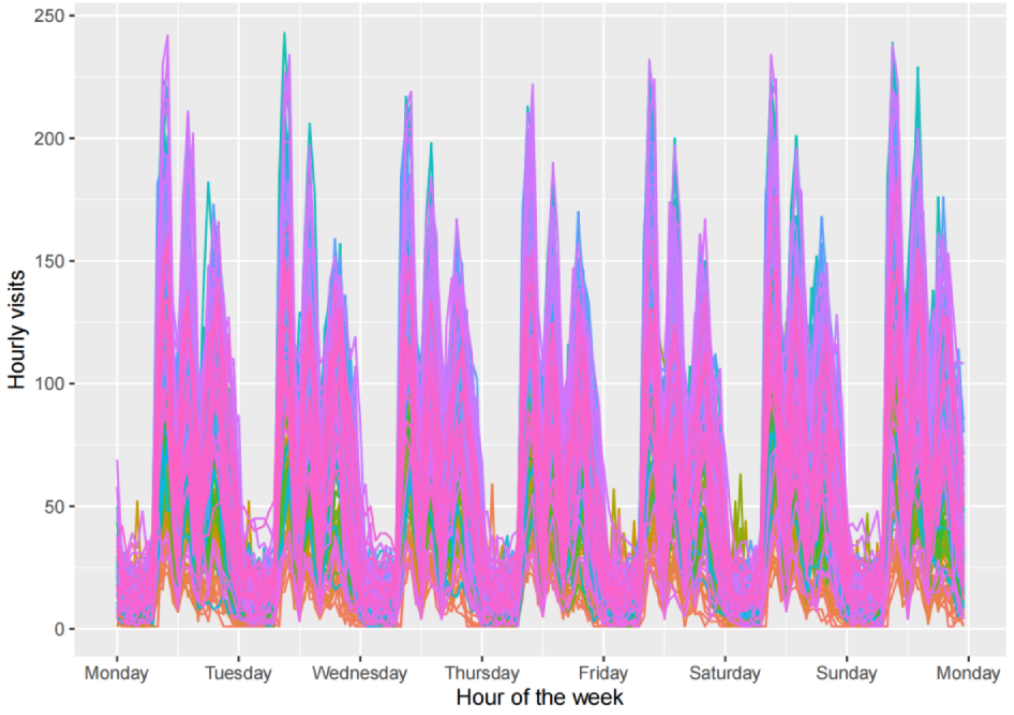
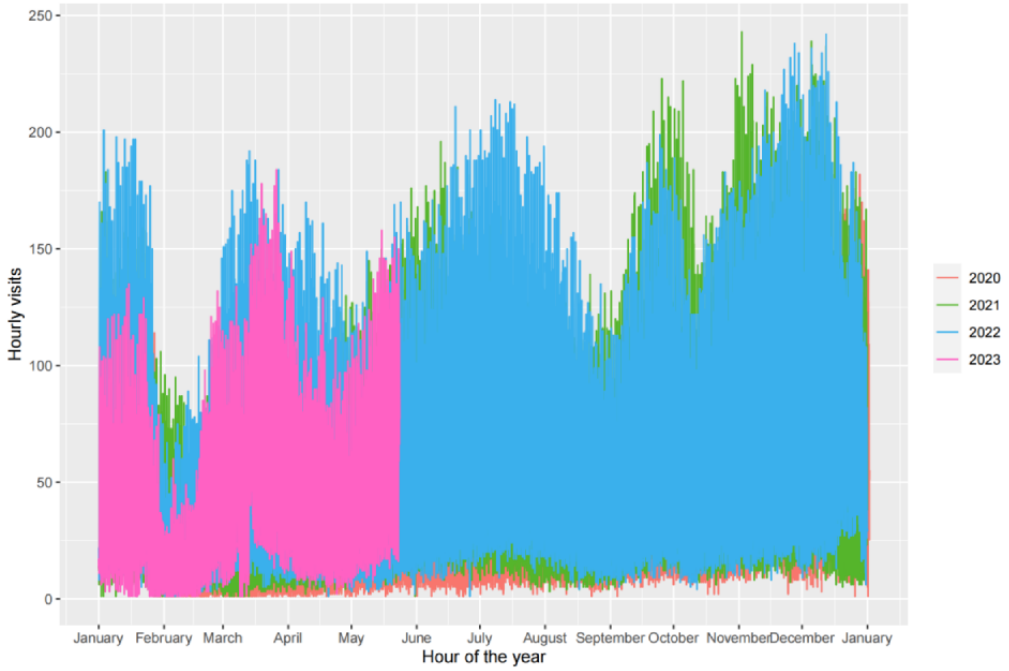


Figure 6. Yearly seasonal patterns in the hourly fever clinic visit time series.



Autocorrelation Function Analysis

The analysis of autocorrelation function (ACF) and partial ACF (PACF) for sample data is a crucial approach for identifying the characteristics of seasonal time series and proposing appropriate candidate models [33]. The lag k autocorrelation coefficient r_k, as measured by ACF, quantifies the linear correlation between 2 observations, y_t and y_{t-k} can be expressed per the following formula:

where y_t represents the value of the time series at time t,  is the mean, and T is the length of the time series. PACF, by contrast, measures the direct correlation between y_t and y_{t-k} while isolating the effects of periods other than k from the analysis.

Figures 7 and 8 depict the sample ACF and PACF for the initial 48 lags derived from the hourly fever clinic visitation data and their corresponding seasonal differential data. As illustrated in Figure 7, the ACF values are uniformly positive and exhibit symmetrical, humped shapes with spike values occurring at multiples of 24-hour intervals, whereas the PACF values exhibit decay in seasonal lags at multiples of 24 hours. These observations suggest that the hourly visit time series is nonstationary and lacks a discernible trend, yet it exhibits strong daily seasonality. Consequently, a seasonal difference by daily period (24 hours) can be applied to generate a stationary time series. As demonstrated in Figure 8, the ACF decays exponentially to approximately 0 after 5 lags, with a downward spike at the first seasonal lag, whereas the PACF exhibits tailing off within each daily periodicity. This indicates the presence of short-term autocorrelation within the differential series as well as a strong negative autocorrelation with 1 seasonal lag. Therefore, it is feasible to achieve short-term forecasts and single seasonal (daily seasonal) forecasts through autocorrelation-based modeling.

Figure 7. Autocorrelation function (ACF) and partial autocorrelation function (PACF) of the hourly fever clinic visit time series.

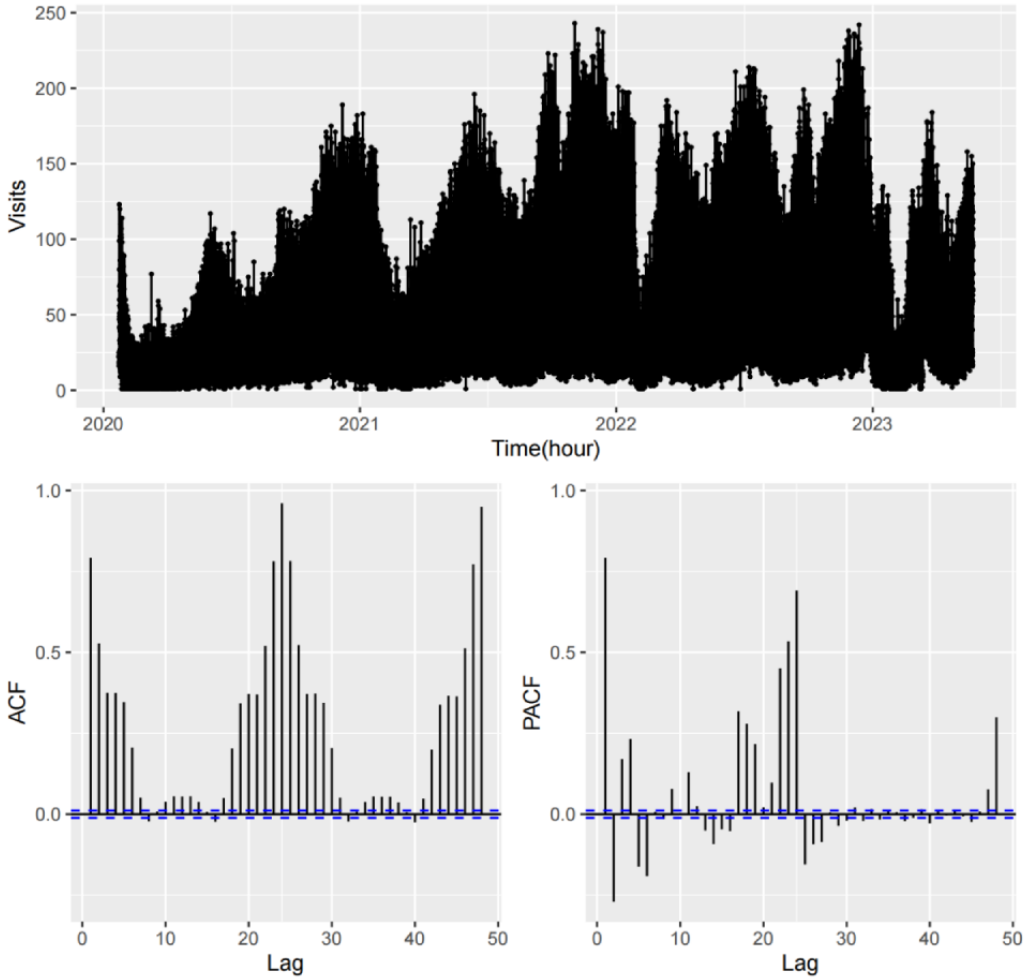
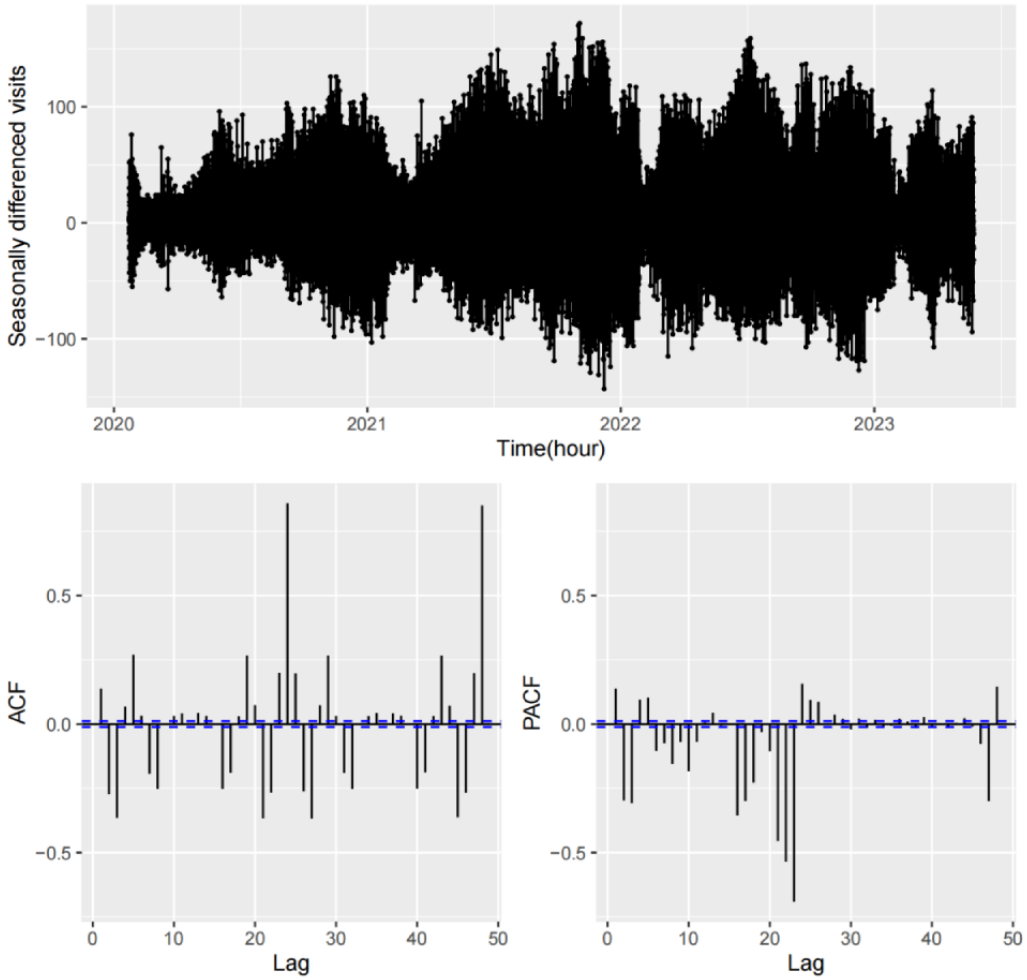


Figure 8. Autocorrelation function (ACF) and partial autocorrelation function (PACF) of the seasonal differential hourly fever clinic visit time series.



STL Analysis

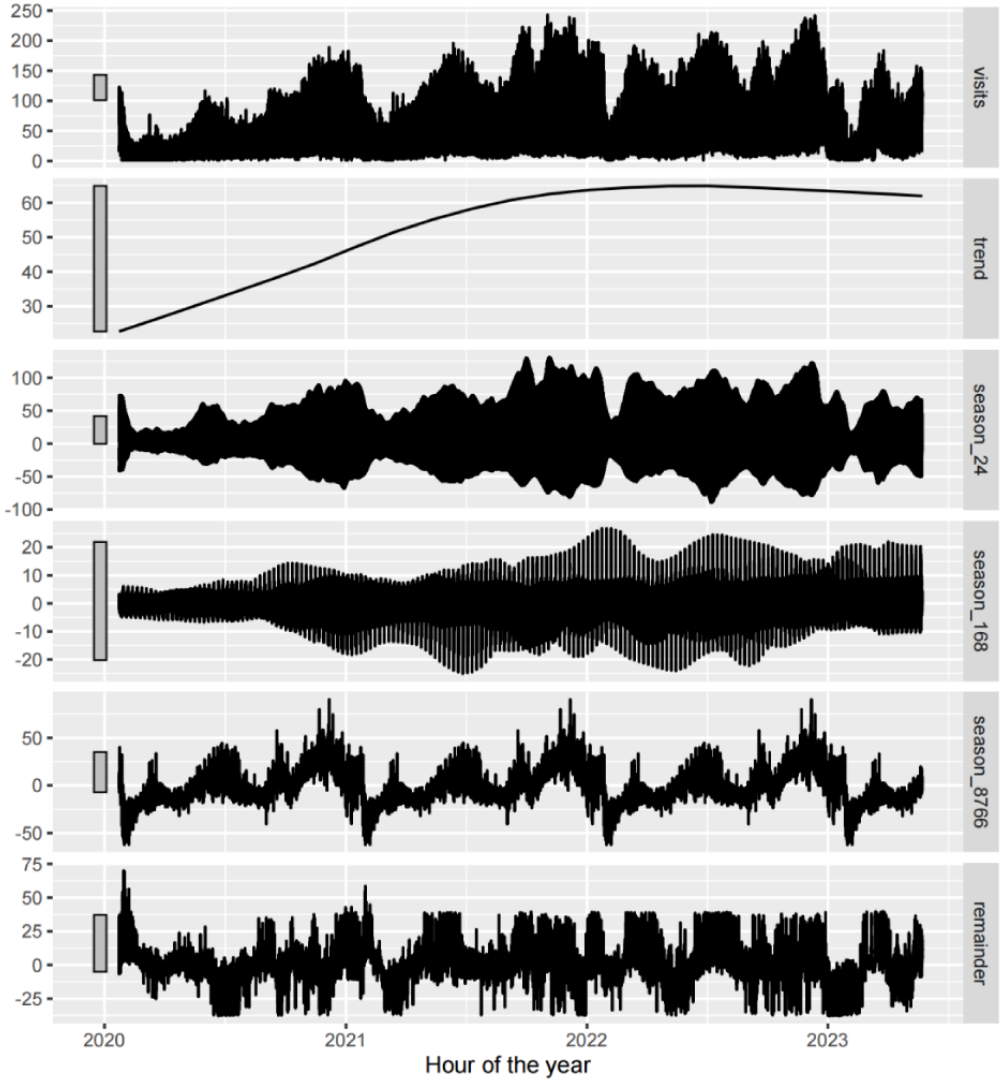
The decomposition of a time series can be used to assess its strength of trend and seasonality. The STL, developed by Cleveland et al [34], constitutes a filtering procedure for decomposing a time series into trend, seasonal, and remainder components in an additive manner. This methodology was subsequently extended to facilitate the decomposition of time series exhibiting multiple seasonal patterns [35].

Figure 9 shows the application of STL to hourly fever clinic visit data, yielding multiple seasonal components as well as trend and remainder components. As expected, 3 seasonal patterns were evident, corresponding to the time of day (third panel), time of week (fourth panel), and time of year (fifth panel).

Note the vertical scales of all panels in Figure 9; the trend panel has the widest bar compared with the other panels, which means that the trend has the narrowest range and, consequently, accounts for only a small proportion of the variation within the data series. The weekly seasonality panel exhibits similar characteristics. In addition, the bars on the daily and yearly seasonality panels are only slightly larger than that on the data panel, indicating that the daily and yearly seasonality signals are largely related to variations in the data series.

It can be inferred that the number of visits exhibited an upward trend during the first 2 years, followed by a modest decline commencing at the end of 2022. However, the impact of this change in trend on the overall time series is negligible. Consequently, the influence of changes in the epidemic containment policy on the data can be ignored. Concurrently, the weekly seasonality inherent within the series is relatively weak and exerts minimal influence on temporal variations within the time series, whereas daily and yearly seasonalities exert a more pronounced effect. Furthermore, it is evident that the daily seasonal pattern undergoes temporal variations, whereas the yearly seasonal pattern appears relatively fixed. Therefore, for long-term forecasting, it is advisable to incorporate yearly seasonal lag data. Finally, the panel at the bottom exhibits large random fluctuations owing to its slightly larger size relative to the data panel. This suggests the presence of additional nonlinear signals infiltrating the residual components, potentially attributable to outliers or unaccountable factors.

Figure 9. Seasonal-trend decomposition using Loess on the hourly fever clinic visit time series.



Theory and Calculations

This section delineates the models used in this study and explains the forecast accuracy measures used in the evaluation of model performance.

Forecasting Models

Before conducting experiments, we established a benchmark method as the standard to ensure that the performance of the other selected models surpasses that of it. As a benchmark approach, it is reasonable to consider the prevailing seasonality in the time series. We set the seasonal naive (SNaive) method as our benchmark, given its utility for highly seasonal data. This method sets each forecast such that it is equal to the last observed value from the corresponding season. Here, the predicted value for the forecast horizon h is considered to be equivalent to the observed value from the previous day. Formally, the forecast for time $T+h$ is expressed as

where m represents the specified seasonal period. This approach requires no parameterization or setup and is frequently used as a benchmark method rather than a model of choice.

Time series exhibiting trends or seasonality are not stationary, as their statistical properties vary over time. In such instances, nonstationary time series can be rendered stationary through the application of differencing techniques. The ARIMA model constitutes a combination of differencing with autoregressive and moving average (MA) components, which were first proposed by Box and Jenkins [36], and is commonly denoted as ARIMA(p,d,q). Autoregressive refers to the regression of a variable in the model on its own lagged or prior values, whereas MA incorporates the dependency between an observation and a residual error derived from an MA model applied to lagged observations. However, the ARIMA model is only applicable to nonseasonal data. The SARIMA model, denoted as ARIMA(p,d,q)(P,D,Q)$_m$, was derived by incorporating additional seasonal terms into the ARIMA model. It should be noted that the SARIMA model can only specify a single seasonal period parameter, rendering it capable of handling only single seasonality. A mathematical exposition of the ARIMA and SARIMA models can be found in Multimedia Appendix 1.

As demonstrated in the *STL analysis* section, STL has been applied to the hourly fever clinic visit time series exhibiting multiple seasonality. This can be conceptualized as decomposition into 3 seasonal components and seasonally

adjusted components. The STL method is based on the performance of weighted local regressions (Loess) on seasonal indices and the trend, with Loess constituting a methodology for estimating nonlinear relationships. This approach confers benefits upon statistical methods by providing a more versatile and robust decomposition procedure than their intrinsic mechanism [37]. The improved algorithm for multiple seasonality uses an iterative method to obtain seasonal components sequentially in ascending order of cycles and finally to compute the trend component in the last iteration [35]. The forecasting approach based on the STL method is defined as the STL forecasting (STLF) model. In this approach, each seasonal component is forecasted by the SNaive method corresponding to the seasonal lags, whereas seasonally adjusted data are forecasted using a nonseasonal ARIMA model. Evidently, this forecasting approach has an advantage in handling multiple seasonal patterns for a long time series, in contrast to the SARIMA model, which is limited to handling single seasonality.

Artificial neural network is based on a structure comprising an input layer, hidden layers, and an output layer, facilitating the modeling of complex nonlinear relationships between response variables and their predictors. The neural network autoregressive (NNAR) model uses lagged values of the time series as inputs to the neural network, with the last observed values from the corresponding season also incorporated as inputs. Generally, the notation NNAR(p,P,K)$_n$ is used to denote the presence of p lag inputs, P seasonal lag inputs, and K neuron nodes in the hidden layer, where m represents the seasonal period. A model can be defined as $y_t=f(y_{t-1},y_{t-2},\ldots,y_{t-p},y_{t-m},\ldots y_{t-Pm})+\varepsilon_t$, where f represents the nonlinear function of the feed-forward network with a single hidden layer, and ε_t is the residual series. In contrast to traditional time-series methods, the network may be applied iteratively, with predictions incorporated as inputs alongside historical data when forecasting additional steps ahead.

In addition, hybrid combinations of the aforementioned models include hybrid SARIMA-STLF, hybrid NNAR-STLF, hybrid SARIMA-NNAR, and hybrid SARIMA-NNAR-STLF. In these hybrid models, multiple forecasts are combined by averaging the forecasts of individual models. This constitutes a straightforward yet effective means of enhancing the forecast accuracy [38]. Moreover, fixed weights facilitate the identification of optimal solutions more effectively than weights that are based on the random statistical variables derived from changing data.

Accuracy Measures

We used root mean squared error (RMSE) and mean absolute error (MAE) to assess the performance of the applied models. Given that the training set is denoted as $\{y_1,y_2,\ldots y_T\}$ and the testing set as $\{y_{T+1},y_{T+2},\ldots\}$, the forecast deviation between the actual observations y_t on the testing set and the corresponding forecasts  can be denoted as  . The formulas for calculating each of these metrics are as follows:

where n represents the total number of observations for the evaluation process. Both RMSE and MAE are scale-dependent metrics. RMSE operates on the principle of averaging errors and is more sensitive to outliers, whereas MAE is calculated from the median of errors and is more robust to outliers.

Results

Overview

In this section, we present the results of the evaluations of all the candidate forecasting models. Hourly visits to the fever clinic were forecasted for the subsequent 15 days, with accuracy calculated based on the forecasted values over this period. The 8 evaluated models were SNaive (as a benchmark), SARIMA, STLF, NNAR, and combinations of these models (excluding SNaive), hybrid SARIMA-STLF, hybrid NNAR-STLF, hybrid SARIMA-NNAR, and hybrid SARIMA-NNAR-STLF. For model training, we used data from January 23, 2020, to April 23, 2023, as the training set and data from April 24, 2023, to May 8, 2023, as the testing set. The performances of the models were compared based on the accuracy of their forecasting results on the testing set.

Model Estimation

Given that the SNaive, SARIMA, and NNAR models are only capable of handling a single seasonal period, we used a daily season period as the seasonal frequency parameter for hourly data. This corresponds to a lag of 24 in SNaive and m=24 in both SARIMA and NNAR. The specifications and estimations for all the models are detailed in Tables 2 and 3. For the selection of hyperparameters and fitting of model parameters, we used the fable toolkit in the R programming language (R Core Team) for implementation.

The SARIMA modeling process was implemented using a variation of the Hyndman-Khandakar algorithm [39], in which parameter D was chosen by an extended Canova-Hansen test [40], and d was chosen through successive Kwiatkowski-Phillips-Schmidt-Shin unit-root tests [41]. Once d and D were determined, a stepwise search was conducted to traverse the ARIMA order space from the initial candidate parameters, selecting values for p, q, P, and Q by minimizing the corrected Akaike's information criterion until the residuals met the white noise conditions. The modeling process for nonseasonal ARIMA was similar, with the exception that the seasonal hyperparameters were set to 0 (P=D=Q=0). After identifying the model orders (hyperparameters p, d, q, P, D, and Q) with the lowest corrected Akaike's information criterion, the model parameters were estimated on the training set using the maximum log-likelihood estimation. In this study, the best-fitting models for SARIMA and STLF on our training data were ARIMA(0,0,5)(0,1,1)$_{24}$ for SARIMA and ARIMA(2,1,2) for STLF. The estimated coefficients for these models are listed in Table 4.

For the NNAR model, the seasonal parameter P was set to 1, and the p parameter was selected from the optimal linear autoregressive(p) model fitted to the seasonally adjusted data (obtained through STL) according to the AIC. The k parameter was rounded to the nearest integer of (p+P+1)/2. In this study,

the best-fitting model for this approach was an average of 20 networks NNAR(44,1,22)$_{24}$, each consisting of a 44×22×1 network with inputs {y_{t-1},y_{t-2},...y_{t-44}} and 22 neurons in the hidden layer. For 1-step-ahead (1 hour ahead) forecasting, available historical inputs were used, whereas for 2-step-ahead forecasting, the 1-step forecast was used as an input along with historical data. This process was executed iteratively until all the required forecasts were computed [38].

Table 2. Summary of the specifications and estimations for different models.

Model and specification	σ^2
SNaive[a]	
SNaive(24)	156.9503
SARIMA[b]	
ARIMA[c](0,0,5)(0,1,1) [24]	103
NNAR[d]	
NNAR(44,1,22) [24]	87.6
STLF[e]	
ARIMA(2,1,2)	55.04
SNaive(24)	1.4266
SNaive(168)	0.2336
SNaive(8766)	0.0045

[a]SNaive: seasonal naive.

[b]SARIMA: seasonal autoregressive integrated moving average.

[c]ARIMA: autoregressive integrated moving average.

[d]NNAR: neural network autoregressive.

[e]STLF: seasonal and trend decomposition using Loess forecasting.

Table 3. Summary of information criterion for autoregressive integrated moving average (ARIMA) models.

Information criterion	ARIMA(0,0,5)(0,1,1) [24]	ARIMA(2,1,2)
AIC[a]	212,731.5	195,029
AICc[b]	212,731.5	195,029
BIC[c]	212,789.3	195,070.2
Log likelihood	–106,358.8	–97,509.48

[a]AIC: Akaike information criterion

[b]AICc: corrected Akaike information criterion.

[c]BIC: Bayesian information criterion.

Table 4. Estimated coefficients of autoregressive integrated moving average (ARIMA) models.

Model and term	Coefficient (SE)	*t*-statistic	*P* value
ARIMA(0,0,5)(0,1,1) [24]			
MA[a](1)	0.3226 (0.0061)	52.49	<.001
MA(2)	0.2177 (0.0063)	34.61	<.001
MA(3)	0.1728 (0.0057)	30.21	<.001
MA(4)	0.1189 (0.0060)	19.78	<.001
MA(5)	0.1088 (0.0058)	18.82	<.001
SMA[b](1)	–0.6929 (0.0046)	–150	<.001
ARIMA(2,1,2)			
AR[c](1)	0.6138 (0.0655)	9.37	<.001
AR(2)	–0.0276 (0.0131)	–2.11	.08
MA(1)	–1.4031 (0.0652)	–21.51	<.001
MA(2)	0.4333 (0.0612)	7.08	.01

[a]MA: moving average.

[b]SMA: seasonal moving average.

[c]AR: autoregressive.

Forecasting Results

The forecasting results of all the evaluated models are shown in Figures 10 and 11. The distribution of the predicted values varies among the models. However, for short-term future data, each model appears to be capable of capturing the majority of their hourly features. It is evident that the prediction and observation lines of all the models are in reasonable agreement. However, as the forecast horizon increases, the agreement between the predicted and actual values of each model diminishes significantly. For further quantitative comparison of the model performance, we calculated the forecast errors within the subsequent 15 days for each model using the accuracy metrics RMSE and MAE. All results are presented in Table 5, which reveals that the hybrid SARIMA-NNAR-STLF model has both the lowest MAE (8.16) and the lowest RMSE (11.09), whereas STLF performs the worst.

Figure 10. Comparison of 15-day forecasts generated using single models with actual observations. NNAR: neural network autoregressive; SARIMA: seasonal autoregressive integrated moving average; SNaive: seasonal naive; STLF: seasonal and trend decomposition using Loess forecasting.

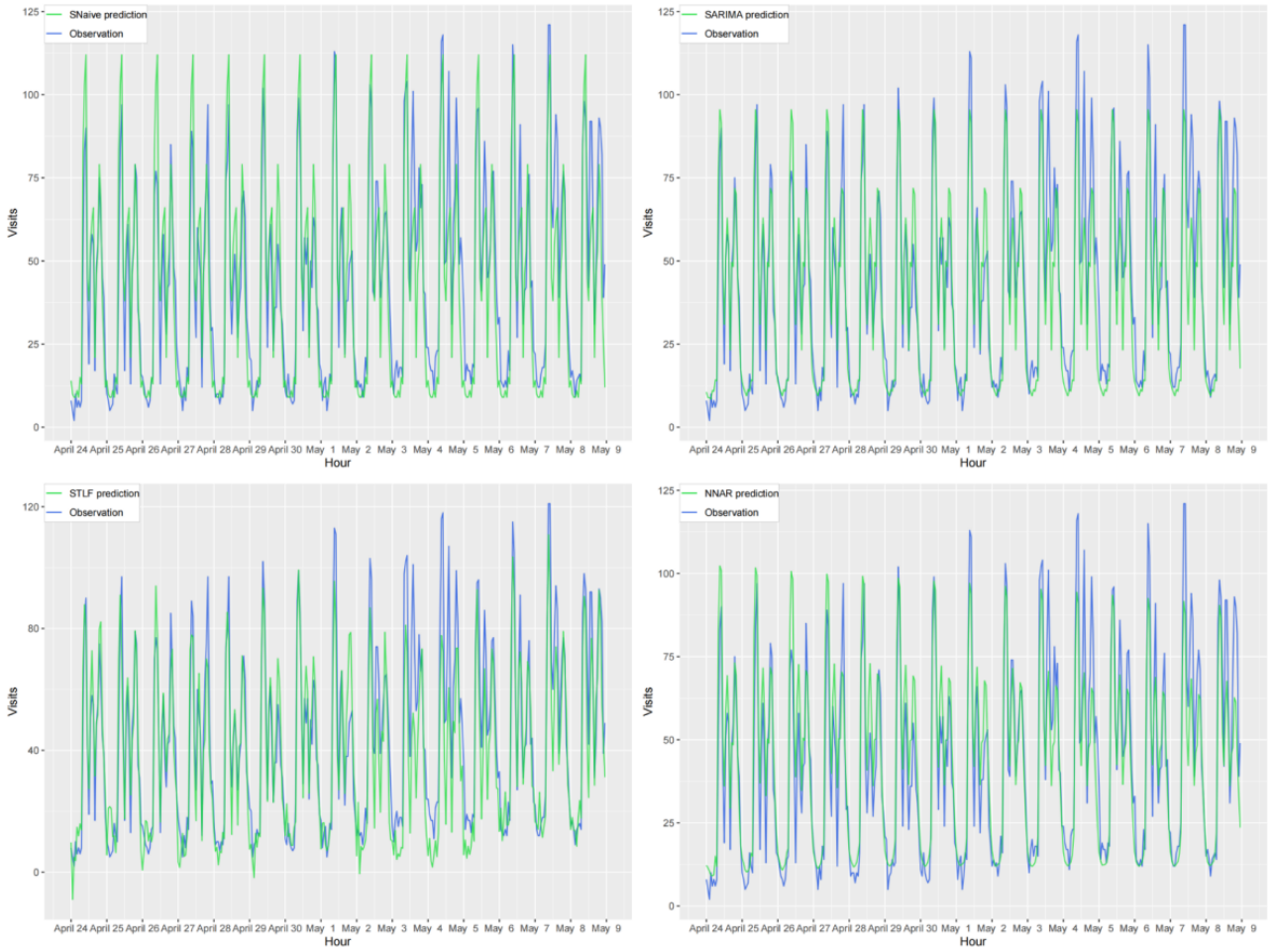


Figure 11. Comparison of 15-day forecasts generated using hybrid models with actual observations. NNAR: neural network autoregressive; SARIMA: seasonal autoregressive integrated moving average; STLF: seasonal and trend decomposition using Loess forecasting.

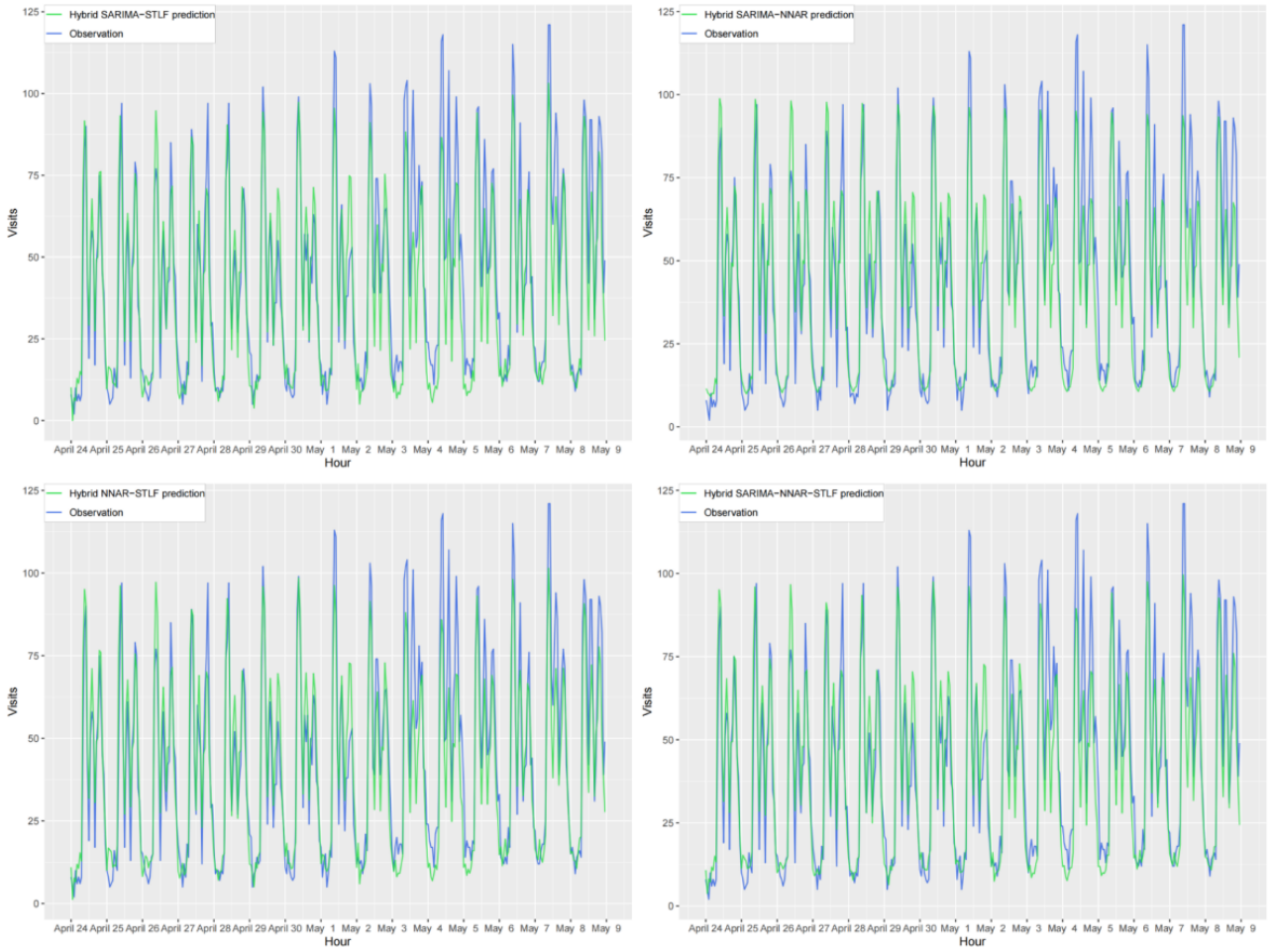


Table 5. Accuracy metrics for all the models on forecasts for the subsequent 15 days (April 24, 2023, to May 8, 2023).

Model	RMSE[a]	MAE[b]
Hybrid NNAR[c]-STLF[d]	11.22	8.38
Hybrid SARIMA[e]-STLF	11.67	8.54
Hybrid SARIMA-NNAR	11.24	8.18
Hybrid SARIMA-NNAR-STLF	11.09	8.16
NNAR	11.65	8.58
SARIMA	11.51	8.33
STLF	13.03	9.69
SNaive[f]	13.02	9.66

[a]RMSE: root mean squared error.

[b]MAE: mean absolute error.

[c]NNAR: neural network autoregressive.

[d]STLF: seasonal and trend decomposition using Loess forecasting.

[e]SARIMA: seasonal autoregressive integrated moving average.

[f]SNaive: seasonal naive.

Cross-Validation

To mitigate the risk of overfitting and determine the optimal hyperparameters for our models, we employed a cross-validation technique based on a rolling forecasting origin [42]. The forecasting origin was advanced incrementally by a fixed number of observations, with forecasts generated at each origin. As the origin progressed, the testing set was incorporated into

the training set for subsequent iterations. Figure 12 illustrates the construction of our cross-validated training and testing sets. To ensure the complete coverage of each month in the testing set, we initiated our analysis with a training set comprising 19,848 observations (827 days), incrementing the size of successive training sets by 720 steps (30 days) with each iteration. This allowed us to generate 1-to-720-step-ahead forecasts and conduct 13 iterations of cross-validation to evaluate the forecasts throughout the entire year. During these iterations, the largest validation set spanned from January 23, 2020, to May 23, 2023, with the training set ranging from January 23, 2020, to April 23, 2023, and the testing set from April 24, 2023, to May 23, 2023.

Forecast accuracy was computed by averaging over the testing sets. This cross-validation method is well suited to account for the temporal dependency between observations in time series, effectively mitigating overfitting while providing a more robust evaluation of model performance [43]. Furthermore, the construction of a rolling length training set is advantageous for validation in multiseasonality cases. The forecast accuracy of all the models for the subsequent 15 days, as determined by cross-validation, is presented in Table 6. The results indicate that the hybrid NNAR-STLF model is the optimal model for this case, as it has both the lowest RMSE (20.1) and the lowest MAE (14.3).

Figure 12. Illustration of the rolling forecasting origin cross-validation methodology.

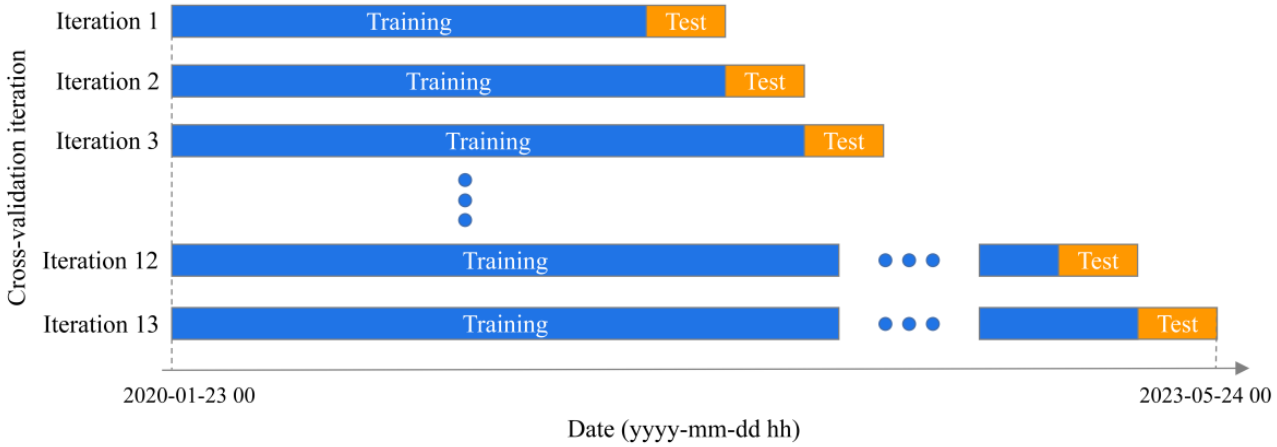


Table 6. Accuracy metrics of all the models for the 15-day forecast by cross-validation.

Model	RMSE[a]	MAE[b]
Hybrid NNAR[c]-STLF[d]	20.1	14.3
Hybrid SARIMA[e]-STLF	21.1	14.6
Hybrid SARIMA-NNAR	22.7	15.5
Hybrid SARIMA-NNAR-STLF	20.4	14.5
NNAR	22.2	15.5
SARIMA	24.1	16.3
STLF	20.8	15.5
SNaive[f]	25.3	17.7

[a]RMSE: root mean squared error.

[b]MAE: mean absolute error.

[c]NNAR: neural network autoregressive.

[d]STLF: seasonal and trend decomposition using Loess forecasting.

[e]SARIMA: seasonal autoregressive integrated moving average.

[f]SNaive: seasonal naive.

Forecast Horizon Accuracy

Although cross-validation has been used in previous studies for performance evaluation, it has rarely been conducted across different forecast horizons. In this study, we not only assessed the accuracy of hourly forecasts for fever clinic visits within a 15-day window but also examined the accuracy of various forecast horizons to provide a more robust basis for identifying the optimal model. Forecast errors across forecast horizons ranging from 1 to 30 days ahead were compared using cross-validation. Specifically, the forecast horizon accuracy was calculated by comparing all the predicted values with their corresponding observed values in the hourly series within a specific 1-day range, representing the model's predictive power

on the i-th day ahead (i=1,...,30). As such, the accuracy metrics for different forecast horizons are independent of one another. The results of our analysis are presented in Figures 13 and 14, which depict the RMSE and MAE values for all the models across different forecast horizons. The calculation of RMSE and MAE was based on the average value obtained across all rolling testing sets in cross-validation.

As shown in Figures 13 and 15, for all the models, the RMSE and MAE values exhibit an overall upward trend as the forecast horizon increases, although not strictly monotonically in some cases. In general, the further ahead we forecast, the greater the uncertainty associated with our predictions. Moreover, the disparity in forecast accuracy among the models also increases with the number of days ahead of the forecast, indicating a widening gap in performance.

Among single models, STLF exhibited superior predictive performance only when forecasting more than 5 days in advance, with the RMSE and MAE values remaining relatively low, especially when forecasting more than 10 days in advance. This highlights the model's advantages in medium- to long-term forecasting. Compared with the benchmark SNaive model, the largest difference in RMSE occurred at a forecast lead time of 17 days, with STLF achieving a value of 22.5 (7.4 lower than that of SNaive), whereas the largest difference in MAE occurred at a forecast lead time of 24 days, with STLF achieving a value of 20.5 (6.2 lower than that of SNaive). In addition, the NNAR model exhibited considerable advantages in short-term forecasting, particularly for forecasts that were 1 to 3 days in advance, with both RMSE and MAE values being the lowest among all the single models (RMSE: 14.3, 15.6, and 16.7; MAE: 10.6, 11.3, and 12.3). However, its performance in medium- to long-term forecasting is more modest.

For hybrid models, the performance was generally superior to that of their constituent single models. Comparing the results for both single and hybrid models, the hybrid NNAR-STLF model exhibited the lowest values for both RMSE and MAE across nearly all forecast horizons, with RMSE ranging from 13.6 (1st day ahead) to 28.3 (30th day ahead) and MAE ranging from 10.8 (1st day ahead) to 21.5 (23rd day ahead). It is evident that the hybrid NNAR-STLF model outperforms in all 3 cases (short-, medium-, and long-term forecasting), indicating that it may be considered the optimal model. However, not all the hybrid models performed well. For example, the hybrid SARIMA-NNAR model exhibited a relatively poor performance compared with both single and other hybrid models. Nonetheless, on the whole, the hybrid approach did have a positive impact on forecasting.

Figure 13. Root mean squared error (RMSE) values calculated for different forecast horizons ranging from 1 to 30 days ahead. NNAR: neural network autoregressive; SARIMA: seasonal autoregressive integrated moving average; SNaive: seasonal naive; STLF: seasonal and trend decomposition using Loess forecasting.

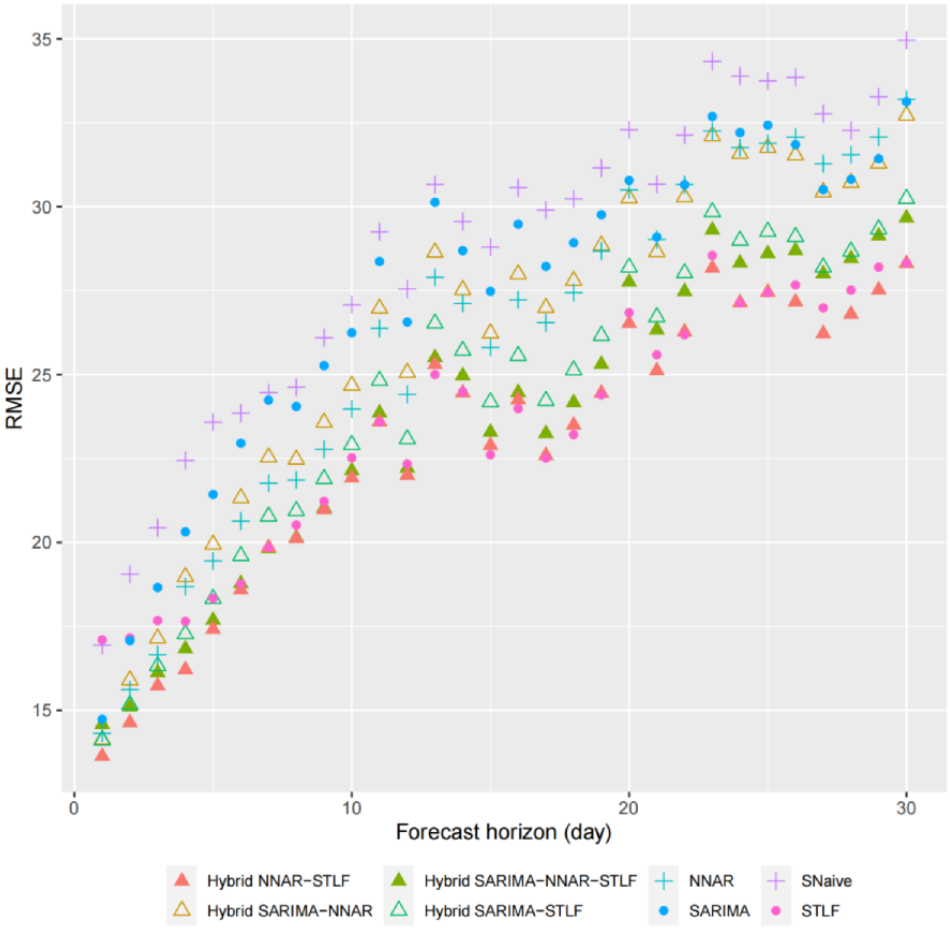


Figure 14. Mean absolute error (MAE) values calculated for different forecast horizons ranging from 1 to 30 days ahead. NNAR: neural network autoregressive; SARIMA: seasonal autoregressive integrated moving average; SNaive: seasonal naive; STLF: seasonal and trend decomposition using Loess forecasting.

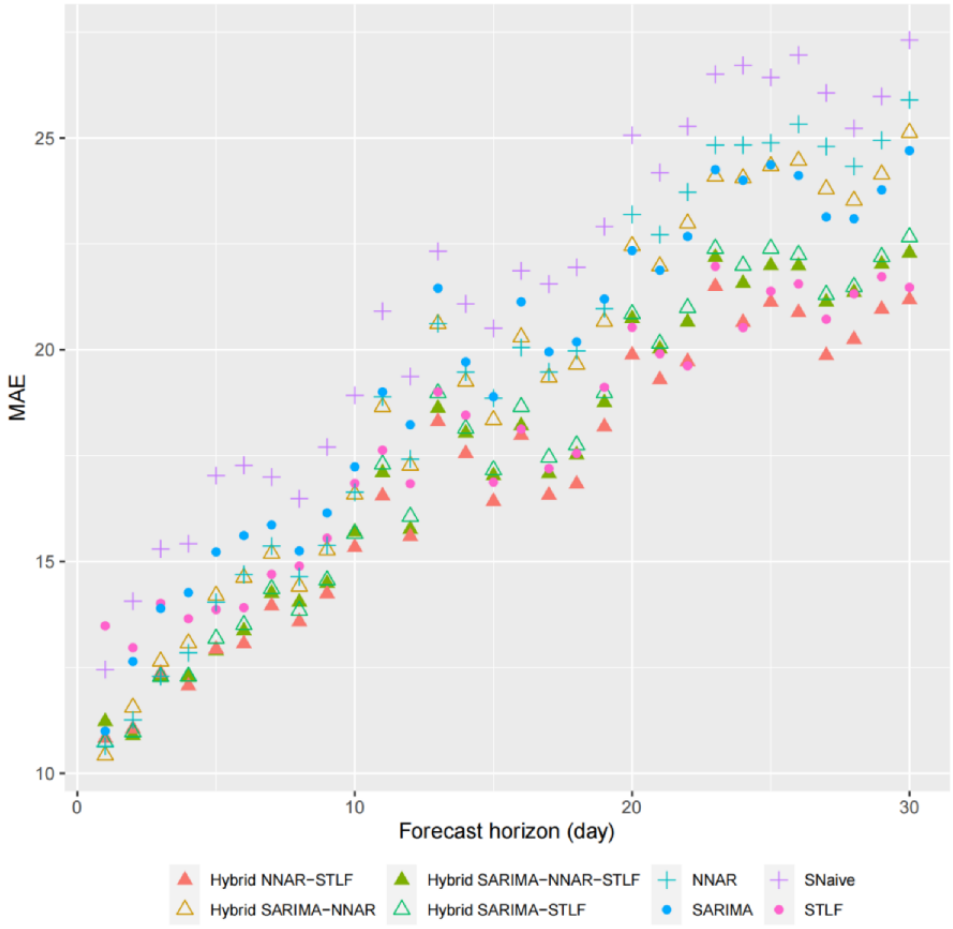
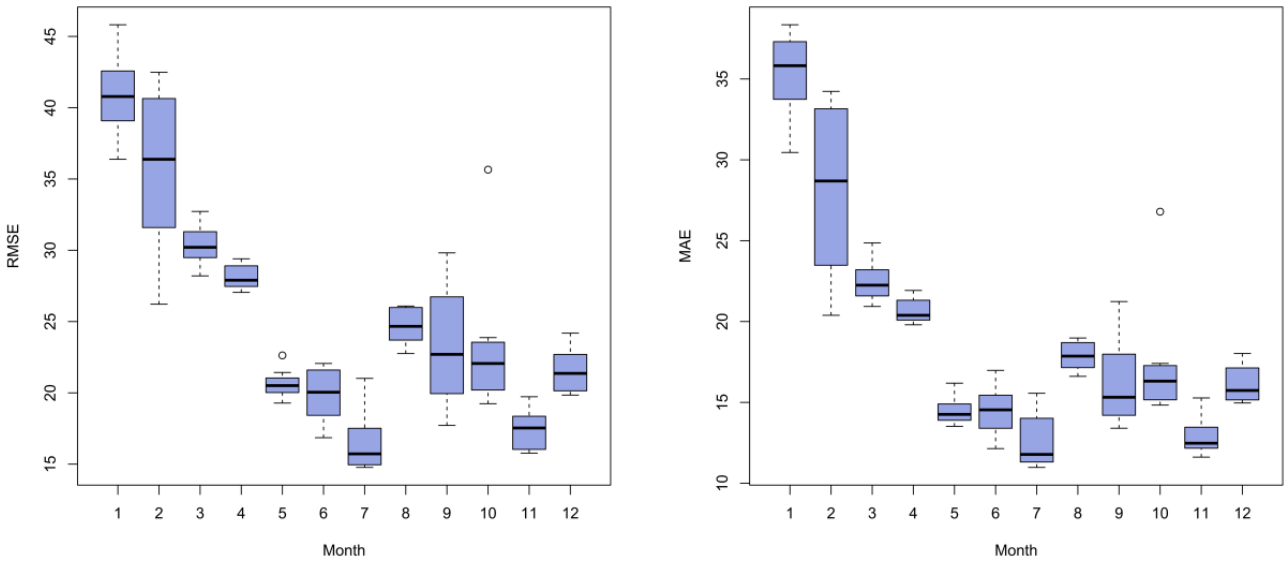


Figure 15. Monthly aggregated root mean squared error (RMSE) and mean absolute error (MAE) box plot.



Discussion

Principal Findings

The hourly visit data from our hospital's fever clinic can be characterized as a time series of long sequences in high frequency. Through exploratory analysis using data visualization, ACF, and STL methods, we observed that the time series exhibited multiseasonality and nonlinearity in its temporal patterns. To achieve our goal of generating hourly forecasts of fever clinic visits in the medium- to long-term horizons, we evaluated an ensemble of individual and hybrid models. In selecting and combining models, we considered their ability to capture multiseasonality and handle nonlinear features.

Our analysis revealed that hybrid models generally outperformed individual models, with the hybrid NNAR-STLF model emerging as the optimal model for our purposes. It exhibited the smallest error in the 15-day forecast horizon (RMSE at 20.1 and MAE at 14.3) and demonstrated a stable advantage in prediction accuracy across forecast horizons ranging from 1 to 30 days ahead. This indicates that it possesses strong scalability and generalization capabilities for predicting multiseasonal periods in time-series data.

Hourly forecasts of fever clinic visits can be leveraged to enhance intelligent outpatient management and provide a sound basis for resource allocation at multiple levels. On the one hand, hospitals can use forecast results to implement flexible scheduling strategies, such as adjusting the number of doctors on duty and modifying registration limitations for specialist doctors and their working hours. On the other hand, during peak seasons and times, hospitals can adapt their facilities and human resources, including the number of service windows, medical technicians, nurses, and outpatient volunteers, to better meet the demands of outpatient operations in accordance with the forecast results. Furthermore, accurate visit forecasts can be used to schedule patient appointments and recommend optimal visitation times, thereby improving efficiency for both hospitals and patients. It is evident that hourly visit forecasts can more effectively support these requirements.

Interpretation of the Findings

Single Models

SNaive and SARIMA are applicable only to time series with a single seasonal pattern. However, the time series for hourly fever clinic visits exhibits 3 seasonal patterns: daily, weekly, and yearly. This results in poor forecast accuracy for both SNaive and SARIMA. By contrast, NNAR takes lagged values as input to the neural network, establishing a more complex nonlinear relationship between forecasts and historical observations than statistical models, allowing it to capture the asymmetry of cycles. However, as it uses forecast values from previous steps as variables for subsequent steps, errors can propagate through the forecast process. In addition, its use of only a single seasonal lag as a forecast variable limits its effectiveness in medium- to long-term forecasting. STLF uses a strategy that captures more comprehensive data characteristics during the forecasting process. It first uses the SNaive method to forecast 3 seasonal components separately and then applies the ARIMA model to forecast seasonally adjusted data. Because hourly data exhibit large daily seasonality variation, relatively low weekly seasonality variation, and relatively fixed yearly seasonality variation, STLF is particularly effective for time series exhibiting yearly seasonal patterns. As such, STLF has an advantage in forecasting long-sequence time series. However, the remainder of STL is somewhat large, indicating that there may be additional factors not accounted for in STLF, such as calendar effects or special events [44].

Hybrid Models

The hybrid model integrating STLF and NNAR has the enhanced ability to capture the multiseasonal and nonlinear characteristics of time series and thus improved forecast accuracy across various forecast horizons. This can explain why the hybrid NNAR-STLF model exhibits the best overall performance among all the models. By contrast, the hybrid SARIMA-NNAR-STLF model exhibits a strengthened autocorrelation between its components and the daily seasonal lag owing to the single-season cycle limitation imposed by the SARIMA component. As a result, the prediction accuracy gap between hybrid NNAR-STLF and hybrid SARIMA-NNAR-STLF widens with increasing forecast horizon, although the inclusion of NNAR does result in a slight improvement in prediction accuracy. The hybrid SARIMA-NNAR model, which lacks the ability to handle multiseasonal patterns, exhibits a distinct disadvantage in medium- to long-term forecasting, with the poorest performance in both cases.

Implication of Errors on Real-World Applications

Inaccurate forecasts can impede effective hospital management and even interfere with decision-making processes. For example, if predicted visits are significantly lower or higher than the actual number, this can result in either inadequate or redundant allocation of personnel, consumables, and facilities. The former can negatively impact the patient's treatment experience through long wait times, overcrowding, and resource shortage, whereas the latter can lead to resource waste and increased operating cost for the hospital.

Limitations and Future Prospects

In this study, we used individual time-series models and their combinations to forecast hourly visits to fever clinics in children's hospitals. Although the optimal hybrid NNAR-STLF model was able to capture the multiseasonality and nonlinear characteristics present in the time-series data, it still produced large errors in forecasting certain months owing to unaccounted external factors. Figure 15 displays the RMSE and MAE box plots for the forecast results from all the models across different months. Statistically, forecast errors were the largest in January and February, with suboptimal errors also observed in March and April. This may be attributed to the influence of the Chinese New Year, which typically falls in January or February. According to Chinese tradition, economic and social activities throughout the community are affected in the months before and after the festival.

Furthermore, the relaxation of COVID-19 containment policies in December of the previous year may have also impacted fever clinic visits. As January was the peak of the first wave of positive COVID-19 cases following the policy adjustment, data after this month lacked sufficient historical training samples, resulting in decreased prediction accuracy after January 2023. Despite this, the continued prevalence of respiratory epidemics in the postepidemic era has led to the retention of high levels of fever clinic visits, because of which fever clinic visit forecasting is still of great significance for clinical decision-making.

In future work, we will account for the effects of moving holidays and disruptive events, as the incorporation of external variables may improve the forecast accuracy [45]. The data set used in this study was obtained from a prominent provincial

public children's hospital in the Yangtze River Delta region of China. Our findings have implications for large hospitals in other regions that have established 24-hour fever clinics. However, to enhance the generalizability of our model, we will incorporate fever clinic visit data from additional medical institutions and construct high-quality, multicenter data sets for model training. Furthermore, this study used a naive averaging strategy to integrate the hybrid model results. The development of more effective fusion strategies represents another important direction for future research.

Conclusions

In this study, we investigated the problem of visit forecasting in a fever clinic in a large public children's hospital in China. Given the changes in clinics' operational mode and patient visitation patterns following the outbreak of the COVID-19 epidemic, developing new forecasting models is essential for supporting intelligent hospital management. The retrospective data on hourly visits to the fever clinic can be characterized as a long-sequence time series in high frequency, with distinct temporal patterns and statistical characteristics inherent to pediatric clinics. Therefore, to identify appropriate models that accurately fit the data and exhibit robust generalization for practical management, we conducted an exploratory data analysis to reveal the seasonality and structural properties of the time-series data. On the basis of these results, we validated an ensemble of time-series models, including individual models and their combinations. We cross-validated their accuracy performance across different forecast horizons. The hybrid NNAR-STLF model was identified as the optimal model for our problem because of its ability to fit multiseasonal patterns and nonlinearity in the time-series data. Its strong performance across different forecast horizons, as indicated by the cross-validation results, further demonstrates its robustness for multiseasonal time series. The model identified in this study is applicable to hospitals with similar outpatient configurations or time series characterized as long sequence in high frequency. We also provided a new research paradigm for other time-series studies, that is, conducting an exploratory analysis revealing data characteristics before model development. However, the existing models do not account for the effects of exogenous variables, such as moving holidays and disruptive events. Future work will explore more comprehensive methods for incorporating external variables and other factors (eg, temperature, humidity, and pollutant levels) into the models to improve their prediction accuracy.

Acknowledgments

This work was supported by the National Key Research and Development (R&D) Program of China (2019YFE0126200), the National Natural Science Foundation of China (62076218), and the Medical Health Science and Technology Project of Zhejiang Provincial Health Commission (2023KY832). The methods of visit forecasting for fever clinics proposed in this work are part of the technical framework designed for the fundings. However, the funding bodies played no role in the writing of the manuscript.

Data Availability

All the data used in this study were obtained from the electronic medical record of the fever clinic at the Children's Hospital of Zhejiang University School of Medicine. The raw electronic medical record data cannot be shared publicly because of the regulatory and legal restrictions imposed by the hospital. However, the data sets generated during this study are available from the corresponding author upon reasonable request, and the code used by this study is public in the GitHub repository [46].

Conflicts of Interest

None declared.

Multimedia Appendix 1
Autoregressive integrated moving average (ARIMA) and seasonal autoregressive integrated moving average (SARIMA) models.
[DOC File , 51 KB - medinform_v11i1e45846_app1.doc]

References

1. Seo JY. Pediatric endocrinology of post-pandemic era. Chonnam Med J 2021 May;57(2):103-107 [FREE Full text] [doi: 10.4068/cmj.2021.57.2.103] [Medline: 34123737]
2. Wang J, Zong L, Zhang J, Sun H, Harold Walline J, Sun P, et al. Identifying the effects of an upgraded 'fever clinic' on COVID-19 control and the workload of emergency department: retrospective study in a tertiary hospital in China. BMJ Open 2020 Aug 20;10(8):e039177 [FREE Full text] [doi: 10.1136/bmjopen-2020-039177] [Medline: 32819955]
3. Bai W, Sha S, Cheung T, Su Z, Jackson T, Xiang YT. Optimizing the dynamic zero-COVID policy in China. Int J Biol Sci 2022 Aug 21;18(14):5314-5316 [FREE Full text] [doi: 10.7150/ijbs.75699] [Medline: 36147473]
4. Notice of the General Office of the National Health Commission on improving the prevention and control of infection in fever clinics and medical institutions. National Health Commission of the People's Republic of China. 2020 Jun 30. URL: http://www.nhc.gov.cn/xcs/zhengcwj/202006/4e456696ceef482996a5bd2c3fb4c3db.shtml [accessed 2021-09-14]

5. Poloniecki JD, Atkinson RW, de Leon AP, Anderson HR. Daily time series for cardiovascular hospital admissions and previous day's air pollution in London, UK. Occup Environ Med 1997 Aug;54(8):535-540 [FREE Full text] [doi: 10.1136/oem.54.8.535] [Medline: 9326156]
6. Atkinson RW, Kang S, Anderson HR, Mills IC, Walton HA. Epidemiological time series studies of PM2.5 and daily mortality and hospital admissions: a systematic review and meta-analysis. Thorax 2014 Jul;69(7):660-665 [FREE Full text] [doi: 10.1136/thoraxjnl-2013-204492] [Medline: 24706041]
7. Atkinson RW, Mills IC, Walton HA, Anderson HR. Fine particle components and health--a systematic review and meta-analysis of epidemiological time series studies of daily mortality and hospital admissions. J Expo Sci Environ Epidemiol 2015 Mar;25(2):208-214 [FREE Full text] [doi: 10.1038/jes.2014.63] [Medline: 25227730]
8. Zheng XY, Ding H, Jiang LN, Chen SW, Zheng JP, Qiu M, et al. Association between air pollutants and asthma emergency room visits and hospital admissions in time series studies: a systematic review and meta-analysis. PLoS One 2015 Sep 18;10(9):e0138146 [FREE Full text] [doi: 10.1371/journal.pone.0138146] [Medline: 26382947]
9. Tian Y, Liu H, Zhao Z, Xiang X, Li M, Juan J, et al. Association between ambient air pollution and daily hospital admissions for ischemic stroke: a nationwide time-series analysis. PLoS Med 2018 Oct 4;15(10):e1002668 [FREE Full text] [doi: 10.1371/journal.pmed.1002668] [Medline: 30286080]
10. van Kasteren ME, Mannien J, Kullberg BJ, de Boer AS, Nagelkerke NJ, Ridderhof M, et al. Quality improvement of surgical prophylaxis in Dutch hospitals: evaluation of a multi-site intervention by time series analysis. J Antimicrob Chemother 2005 Dec;56(6):1094-1102. [doi: 10.1093/jac/dki374] [Medline: 16234334]
11. Ordon M, Urbach D, Mamdani M, Saskin R, D'A Honey RJ, Pace KT. The surgical management of kidney stone disease: a population based time series analysis. J Urol 2014 Nov;192(5):1450-1456. [doi: 10.1016/j.juro.2014.05.095] [Medline: 24866599]
12. Sivasubramaniam V, Patel HC, Ozdemir BA, Papadopoulos MC. Trends in hospital admissions and surgical procedures for degenerative lumbar spine disease in England: a 15-year time-series study. BMJ Open 2015 Dec 15;5(12):e009011 [FREE Full text] [doi: 10.1136/bmjopen-2015-009011] [Medline: 26671956]
13. Sun J, Lin Q, Zhao P, Zhang Q, Xu K, Chen H, et al. Reducing waiting time and raising outpatient satisfaction in a Chinese public tertiary general hospital-an interrupted time series study. BMC Public Health 2017 Aug 22;17(1):668 [FREE Full text] [doi: 10.1186/s12889-017-4667-z] [Medline: 28830400]
14. Mulholland RH, Wood R, Stagg HR, Fischbacher C, Villacampa J, Simpson CR, et al. Impact of COVID-19 on accident and emergency attendances and emergency and planned hospital admissions in Scotland: an interrupted time-series analysis. J R Soc Med 2020 Nov;113(11):444-453 [FREE Full text] [doi: 10.1177/0141076820962447] [Medline: 33012218]
15. Wagner AK, Soumerai SB, Zhang F, Ross-Degnan D. Segmented regression analysis of interrupted time series studies in medication use research. J Clin Pharm Ther 2002 Aug;27(4):299-309. [doi: 10.1046/j.1365-2710.2002.00430.x] [Medline: 12174032]
16. Ansari F, Gray K, Nathwani D, Phillips G, Ogston S, Ramsay C, et al. Outcomes of an intervention to improve hospital antibiotic prescribing: interrupted time series with segmented regression analysis. J Antimicrob Chemother 2003 Nov;52(5):842-848. [doi: 10.1093/jac/dkg459] [Medline: 14563900]
17. Mol PG, Wieringa JE, Nannanpanday PV, Gans RO, Degener JE, Laseur M, et al. Improving compliance with hospital antibiotic guidelines: a time-series intervention analysis. J Antimicrob Chemother 2005 Apr;55(4):550-557 [FREE Full text] [doi: 10.1093/jac/dki037] [Medline: 15728141]
18. Willemsen I, Cooper B, van Buitenen C, Winters M, Andriesse G, Kluytmans J. Improving quinolone use in hospitals by using a bundle of interventions in an interrupted time series analysis. Antimicrob Agents Chemother 2010 Sep;54(9):3763-3769 [FREE Full text] [doi: 10.1128/AAC.01581-09] [Medline: 20585135]
19. Hertzum M. Forecasting hourly patient visits in the emergency department to counteract crowding. Open Ergonomics J 2017 Mar 31;10(1):1-13 [FREE Full text] [doi: 10.2174/1875934301710010001]
20. Choudhury A, Urena E. Forecasting hourly emergency department arrival using time series analysis. British Journal of Healthcare Management 2020 Jan 02;26(1):34-43 [FREE Full text] [doi: 10.12968/bjhc.2019.0067]
21. Becerra M, Jerez A, Aballay B, Garcés HO, Fuentes A. Forecasting emergency admissions due to respiratory diseases in high variability scenarios using time series: a case study in Chile. Sci Total Environ 2020 Mar 01;706:134978. [doi: 10.1016/j.scitotenv.2019.134978] [Medline: 31862585]
22. Cheng Q, Argon NT, Evans CS, Liu Y, Platts-Mills TF, Ziya S. Forecasting emergency department hourly occupancy using time series analysis. Am J Emerg Med 2021 Oct;48:177-182. [doi: 10.1016/j.ajem.2021.04.075] [Medline: 33964692]
23. Whitt W, Zhang X. Forecasting arrivals and occupancy levels in an emergency department. Oper Res Health Care 2019 Jun;21:1-18 [FREE Full text] [doi: 10.1016/j.orhc.2019.01.002]
24. Harrou F, Dairi A, Kadri F, Sun Y. Forecasting emergency department overcrowding: a deep learning framework. Chaos Solitons Fractals 2020 Oct;139:110247 [FREE Full text] [doi: 10.1016/j.chaos.2020.110247] [Medline: 32982079]
25. Etu EE, Monplaisir L, Aguwa C, Arslanturk S, Masoud S, Krupp S, et al. 33 forecasting daily patient arrivals during COVID-19 in emergency departments: a deep learning approach. Ann Emerg Med 2021 Oct;78(4):S14 [FREE Full text] [doi: 10.1016/j.annemergmed.2021.09.041]

26. Zhang Y, Zhang J, Tao M, Shu J, Zhu D. Forecasting patient arrivals at emergency department using calendar and meteorological information. Appl Intell (Dordr) 2022;52(10):11232-11243 [FREE Full text] [doi: 10.1007/s10489-021-03085-9] [Medline: 35079202]
27. Sudarshan VK, Brabrand M, Range TM, Wiil UK. Performance evaluation of emergency department patient arrivals forecasting models by including meteorological and calendar information: a comparative study. Comput Biol Med 2021 Aug;135:104541. [doi: 10.1016/j.compbiomed.2021.104541] [Medline: 34166880]
28. Khaldi R, Afia AE, Chiheb R. Forecasting of weekly patient visits to emergency department: real case study. Procedia Comput Sci 2019;148:532-541 [FREE Full text] [doi: 10.1016/j.procs.2019.01.026]
29. Deng Y, Fan H, Wu S. A hybrid ARIMA-LSTM model optimized by BP in the forecast of outpatient visits. J Ambient Intell Humaniz Comput 2020 Oct 19;14(5):5517-5527 [FREE Full text] [doi: 10.1007/s12652-020-02602-x]
30. Perone G. Comparison of ARIMA, ETS, NNAR, TBATS and hybrid models to forecast the second wave of COVID-19 hospitalizations in Italy. Eur J Health Econ 2022 Aug 04;23(6):917-940 [FREE Full text] [doi: 10.1007/s10198-021-01347-4] [Medline: 34347175]
31. Population status of children in China in 2015: facts and figures. The United Nations International Children's Emergency Fund, China. URL: https://www.unicef.cn/en/reports/population-status-children-china-2015 [accessed 2022-12-30]
32. International statistical classification of diseases and related health problems 10th revision. World Health Organization. 2020. URL: https://icd.who.int/browse10/2019/en [accessed 2020-12-30]
33. Hamilton DC, Watts DG. Interpreting partial autocorrelation functions of seasonal time series models. Biometrika 1978;65(1):135-140 [FREE Full text] [doi: 10.1093/biomet/65.1.135]
34. Cleveland RB, Cleveland WS, McRae JE, Terpenning I. STL: a seasonal-trend decomposition procedure based on loess. J Off Stat 1990;6(1):3-73 [FREE Full text]
35. Bandara K, Hyndman RJ, Bergmeir C. MSTL: a seasonal-trend decomposition algorithm for time series with multiple seasonal patterns. arXiv Preprint posted online July 28, 2021 [FREE Full text] [doi: 10.1504/ijor.2022.10048281]
36. Box GE, Jenkins GM. Time series analysis: forecasting and control. Holden-Day. 1970. URL: http://garfield.library.upenn.edu/classics1989/A1989AV48500001.pdf [accessed 2022-09-14]
37. Ouyang Z, Ravier P, Jabloun M. STL decomposition of time series can benefit forecasting done by statistical methods but not by machine learning ones. Eng Proc 2021;5(1):42 [FREE Full text] [doi: 10.3390/engproc2021005042]
38. Hyndman RJ, Athanasopoulos G. Forecasting: principles and practice, 3rd edition. Monash University, Australia. URL: https://otexts.com/fpp3/ [accessed 2022-12-30]
39. Hyndman RJ, Khandakar Y. Automatic time series forecasting: the forecast package for R. J Stat Soft 2008;27(3):1-22 [FREE Full text] [doi: 10.18637/jss.v027.i03]
40. Canova F, Hansen BE. Are seasonal patterns constant over time? A test for seasonal stability. J Bus Econ Stat 1995;13(3):237-252 [FREE Full text] [doi: 10.1080/07350015.1995.10524598]
41. Kwiatkowski D, Phillips PC, Schmidt P, Shin Y. Testing the null hypothesis of stationarity against the alternative of a unit root: how sure are we that economic time series have a unit root? J Econom 1992 Oct;54(1-3):159-178 [FREE Full text] [doi: 10.1016/0304-4076(92)90104-y]
42. Tashman LJ. Out-of-sample tests of forecasting accuracy: an analysis and review. Int J Forecast 2000 Oct;16(4):437-450 [FREE Full text] [doi: 10.1016/s0169-2070(00)00065-0]
43. Bergmeir C, Benítez JM. On the use of cross-validation for time series predictor evaluation. Inf Sci 2012 May 15;191:192-213 [FREE Full text] [doi: 10.1016/j.ins.2011.12.028]
44. Trull O, García-Díaz JC, Peiró-Signes A. Multiple seasonal STL decomposition with discrete-interval moving seasonalities. Appl Math Comput 2022 Nov;433:127398 [FREE Full text] [doi: 10.1016/j.amc.2022.127398]
45. Gao J, Zhang P. China's public health policies in response to COVID-19: from an "authoritarian" perspective. Front Public Health 2021 Dec 15;9:756677 [FREE Full text] [doi: 10.3389/fpubh.2021.756677] [Medline: 34976920]
46. Experimental code for JMIR manuscript. GitHub. URL: https://github.com/cutezw/JMIR_45846 [accessed 2023-09-11]

Abbreviations

ACF: autocorrelation function
ARIMA: autoregressive integrated moving average
EMR: electronic medical record
LSTM: long short-term memory
MA: moving average
MAE: mean absolute error
NNAR: neural network autoregressive
PACF: partial autocorrelation function
RMSE: root mean squared error
SARIMA: seasonal autoregressive integrated moving average
SNaive: seasonal naive

STL: seasonal-trend decomposition using Loess
STLF: seasonal and trend decomposition using Loess forecasting

Edited by C Lovis; submitted 19.01.23; peer-reviewed by Z Zhang, J Zhang; comments to author 15.06.23; revised version received 20.07.23; accepted 10.08.23; published 20.09.23.

Please cite as:
Zhang W, Zhu Z, Zhao Y, Li Z, Chen L, Huang J, Li J, Yu G
Analyzing and Forecasting Pediatric Fever Clinic Visits in High Frequency Using Ensemble Time-Series Methods After the COVID-19 Pandemic in Hangzhou, China: Retrospective Study
JMIR Med Inform 2023;11:e45846
URL: https://medinform.jmir.org/2023/1/e45846
doi:10.2196/45846
PMID:37728972

Original Paper

Synthetic Tabular Data Based on Generative Adversarial Networks in Health Care: Generation and Validation Using the Divide-and-Conquer Strategy

Ha Ye Jin Kang[1,2*], MS; Erdenebileg Batbaatar[3*], PhD; Dong-Woo Choi[3], PhD; Kui Son Choi[3,4], PhD, Prof Dr; Minsam Ko[1,5], PhD, Prof Dr; Kwang Sun Ryu[2,3], PhD, Prof Dr

[1]Department of Applied Artificial Intelligence, Hanyang University, Ansan, Republic of Korea

[2]Department of Cancer AI & Digital Health, Graduate School of Cancer Science and Policy, National Cancer Center, Gyeonggi-do, Republic of Korea

[3]National Cancer Data Center, National Cancer Control Institute, National Cancer Center, Gyeonggi-do, Republic of Korea

[4]Department of Cancer Control and Policy, Graduate School of Cancer Science and Policy, National Cancer Center, Gyeonggi-do, Republic of Korea

[5]Department of Human-Computer Interaction, Hanyang University, Ansan, Republic of Korea

[*]these authors contributed equally

Corresponding Author:
Kwang Sun Ryu, PhD, Prof Dr
National Cancer Data Center
National Cancer Control Institute
National Cancer Center
323 Ilsan-ro, Ilsandong-gu, Goyang-si
Gyeonggi-do, 10408
Republic of Korea
Phone: 82 31 920 0652
Email: niceplay13@ncc.re.kr

Abstract

Background: Synthetic data generation (SDG) based on generative adversarial networks (GANs) is used in health care, but research on preserving data with logical relationships with synthetic tabular data (STD) remains challenging. Filtering methods for SDG can lead to the loss of important information.

Objective: This study proposed a divide-and-conquer (DC) method to generate STD based on the GAN algorithm, while preserving data with logical relationships.

Methods: The proposed method was evaluated on data from the Korea Association for Lung Cancer Registry (KALC-R) and 2 benchmark data sets (breast cancer and diabetes). The DC-based SDG strategy comprises 3 steps: (1) We used 2 different partitioning methods (the class-specific criterion distinguished between survival and death groups, while the Cramer V criterion identified the highest correlation between columns in the original data); (2) the entire data set was divided into a number of subsets, which were then used as input for the conditional tabular generative adversarial network and the copula generative adversarial network to generate synthetic data; and (3) the generated synthetic data were consolidated into a single entity. For validation, we compared DC-based SDG and conditional sampling (CS)–based SDG through the performances of machine learning models. In addition, we generated imbalanced and balanced synthetic data for each of the 3 data sets and compared their performance using 4 classifiers: decision tree (DT), random forest (RF), Extreme Gradient Boosting (XGBoost), and light gradient-boosting machine (LGBM) models.

Results: The synthetic data of the 3 diseases (non–small cell lung cancer [NSCLC], breast cancer, and diabetes) generated by our proposed model outperformed the 4 classifiers (DT, RF, XGBoost, and LGBM). The CS- versus DC-based model performances were compared using the mean area under the curve (SD) values: 74.87 (SD 0.77) versus 63.87 (SD 2.02) for NSCLC, 73.31 (SD 1.11) versus 67.96 (SD 2.15) for breast cancer, and 61.57 (SD 0.09) versus 60.08 (SD 0.17) for diabetes (DT); 85.61 (SD 0.29) versus 79.01 (SD 1.20) for NSCLC, 78.05 (SD 1.59) versus 73.48 (SD 4.73) for breast cancer, and 59.98 (SD 0.24) versus 58.55 (SD 0.17) for diabetes (RF); 85.20 (SD 0.82) versus 76.42 (SD 0.93) for NSCLC, 77.86 (SD 2.27) versus 68.32 (SD 2.37) for breast cancer, and 60.18 (SD 0.20) versus 58.98 (SD 0.29) for diabetes (XGBoost); and 85.14 (SD 0.77) versus 77.62 (SD 1.85) for NSCLC, 78.16 (SD 1.52) versus 70.02 (SD 2.17) for breast cancer, and 61.75 (SD 0.13) versus 61.12 (SD 0.23) for diabetes (LGBM). In addition, we found that balanced synthetic data performed better.

Conclusions: This study is the first attempt to generate and validate STD based on a DC approach and shows improved performance using STD. The necessity for balanced SDG was also demonstrated.

(JMIR Med Inform 2023;11:e47859) doi:10.2196/47859

KEYWORDS

generative adversarial networks; GAN; synthetic data generation; synthetic tabular data; lung cancer; machine learning; mortality prediction

Introduction

Machine learning (ML) techniques have been applied in health care with remarkable success over the past decade. ML has the potential to improve tasks in various fields in the medical industry [1]. Analysis of clinical data to predict risk factors and degrees of association between diseases [2] is one of the major advancements achieved using ML. However, the application of ML in real-world clinical environments remains difficult owing to clinical limitations, such as data scarcity, data privacy, and data imbalance [3]. In this context, generative adversarial networks (GANs) [4] have emerged as one of the most important types of ML-based generative models in health care [5].

GAN algorithms generate large amounts of synthetic patient data, which can serve as an appropriate alternative to real data [6-8]. A GAN comprises 2 models trained using an adversarial process, in which one model—the "generator"—generates synthetic data, while the other—the "discriminator"—distinguishes between real and synthetic data. Conventional GAN algorithms have been enhanced and repurposed for clinical tabular data [9-11]. In addition, GANs alleviate clinical limitations and facilitate the application of ML in health care [3,12]. Beaulieu-Jones et al [13] used the auxiliary classifier generative adversarial network (ACGAN) to generate synthetic SPRINT (Systolic Blood Pressure Intervention Trial) data for privacy-preserving data sharing. Baowaly et al [14] generated synthetic electronic health record data using the medical generative adversarial network (MedBGAN) to resolve the data-sharing problem. Izonin et al [15] created an enlarged data set based on a GAN to improve the accuracy of diagnostics tasks. Wang et al [16] developed a framework to generate and evaluate synthetic data, while simultaneously preserving the complexities of real data and ensuring privacy.

Nevertheless, the application of existing models and algorithms, which are not tailor-made for tabular health care data, to synthetic data generation (SDG) in this field remains unsuitable. Some do not consider the characteristics of health care tabular data [17]. To generate synthetic tabular data (STD), while preserving data with logical relationships, the relationships between columns in the original data (OD) must be considered. The OD have a logical relationship between each column: For example, measurement of the drinking attribute is performed using the binary classification "yes" or "no." If the value of this attribute is "no" in some records, the corresponding value of the subattribute "How much do you drink per week?" must be 0. However, poorly designed GANs may generate synthetic data containing impossible values, for example, a record indicating "drinking: no" and "How much do you drink per week?: 10." This can potentially affect the quality of the generated synthetic data and make them unreliable for certain analyses. To prevent this, filtering methods in GANs have been developed. Both the conditional tabular generative adversarial network (CTGAN) [18] and the copula generative adversarial network (CopulaGAN) [19] use conditional sampling (CS) as a filtering method to forcibly express logical relationships. CS is a method used in the CTGAN and CopulaGAN. CS works through a process of rejection sampling, in which multiple iterations of row sampling occur until a satisfactory row that meets the established conditions is obtained. The performance is also compared on balanced and imbalanced synthetic data sets. However, filtering methods exclude record data based on predefined condition columns after STD generation, ignoring meaningful information contained in the excluded records. To mitigate this risk, it is important to carefully consider the specified conditions and to ensure that they are representative of the broader population.

In addition, although it is generally accepted that balanced data perform better in classification, there has been little research based on experiments that clearly demonstrate how much class-balanced tabular synthetic data are required to improve model performance. Therefore, our experiments suggest that when creating a reference table, we should consider how much data to create so that the classes are balanced when generating synthetic data.

In this study, we proposed an SDG framework to overcome the aforementioned challenges in clinical data generation. The remainder of the paper is organized as follows. The *Methods* section describes the basic characteristics of the study population and the divide-and-conquer (DC)–based SDG strategy, defines the division criteria, presents the problem statement for a filtering method, and presents the SDG process and verification methods. The *Results* section compares the prediction performances of the proposed approach and CS. Moreover, the quality of the generated STD is estimated. Finally, the *Discussion* section elaborates further on the experimental design, results, limitations, and conclusions.

Methods

Ethical Considerations

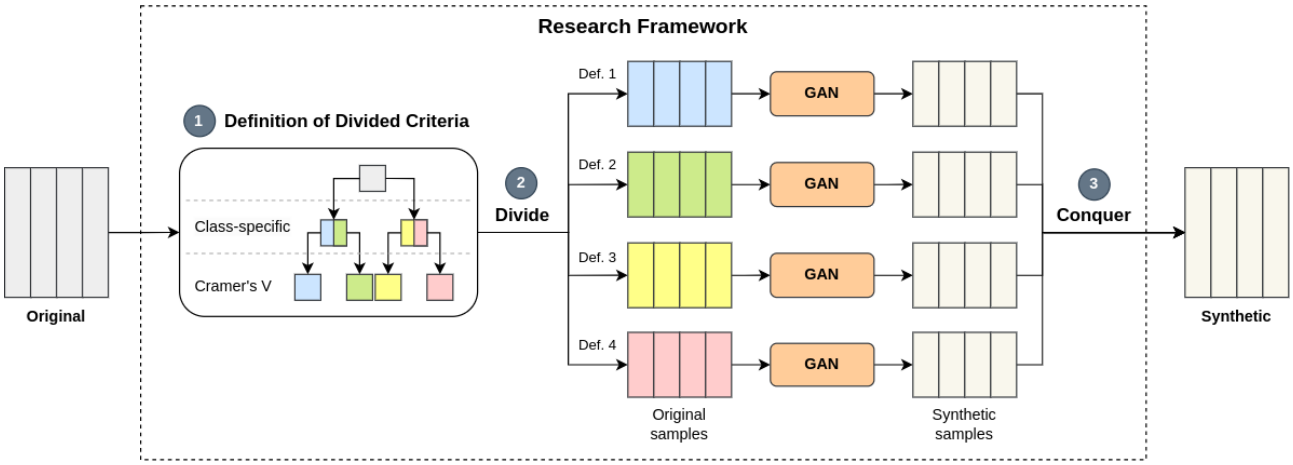
The study design was approved by the Ethics Review Board of the National Cancer Center Institutional Review Board (IRB no: NCC2022-0281).

Research Framework

The DC-based research framework, as depicted in Figure 1, generates STD, while preserving data with logical relationships to enable comparison in terms of data reliability and

investigation of the factors affecting ML model performance. In the division step, the entire data set was divided into several subsets based on the division criteria. In the conquer step, different subsets generated via the GAN were combined into 1. Following STD generation, model performances achieved using classification algorithms, such as the decision tree (DT), the random forest (RF), Extreme Gradient Boosting (XGBoost), and the light gradient-boosting machine (LGBM), in both DC- and CS-based approaches of the CTGAN and CopulaGAN were compared. Moreover, ML model performance on balanced synthetic data and imbalanced synthetic data was also compared.

Figure 1. DC-based research framework using GANs. Def.: definition; GAN: generative adversarial network.



Definition of the Cramer V and Class-Specific Division Criteria

In this study, the division criteria involved 2 main components, class-specific and the Cramer V criteria. The class-specific criterion enabled the selection of different feature subsets for all classes, allowing for subsamples that were tailored to the unique characteristics and behavior of each class. Meanwhile, the high correlation-based criterion identified variables with high correlation scores by computing the Cramer V correlation matrix. The reason for using the Cramer V criterion as the second division criterion was to preserve the logical relationships in the OD in the synthetic data. These variables were then used as the basis for defining logical relationships of data that would guide the division of the data set into subsamples. These 2 criteria provided a robust and effective approach to analyzing the data and identifying meaningful patterns and relationships within them.

Class-Specific Criterion

We used a class-specific division criterion [20], which enabled the selection of different feature subsets for all classes. This yielded a comprehensive list of data set allocation attributes and values by deconstructing the OD set into smaller, more refined subsets. These subsets were subsequently classified based on the dependent classes between different classes, which in turn represented unique sets of class-based criteria. This enabled the selection of a different feature subset for each class. This approach is particularly useful when dealing with data sets comprising classes with unique characteristics and behaviors. The class-specific criterion enables the creation of subsamples tailored to each class, in turn leading to more accurate predictions and better insights.

Cramer V Criterion

We used the Cramer V correlation to identify high correlation patterns in the data set. The Cramer V criterion is a measure of substantive importance used to measure how strongly 2 categorical fields are related to each other. Its value ranges from 0 to 1, where 0 represents no association between the categorical variables and 1 represents complete association in the contingency table. The Cramer V correlation coefficient can be calculated using the formula provided in Equation 1:

where V denotes the Cramer V correlation coefficient, χ^2 denotes the chi-square statistic of the contingency table, N denotes the total number of observations in the contingency table, r denotes the number of rows in the contingency table, and c denotes the number of columns in the contingency table.

The steps in calculating the Cramer V correlation coefficient are as follows:

- Step 1: Calculate χ^2, which is a measure of the association or independence between 2 categorical variables represented in a contingency table and quantifies how much the observed frequencies would deviate from the expected frequencies if the variables were independent. A higher χ^2 value suggests a stronger association between the variables.
- Step 2: Determine the scaling factor, which is necessary to normalize the Cramer V correlation coefficient. The scaling factor is calculated as min(r – 1, c – 1), where (r – 1) and (c – 1) represent the degrees of freedom associated with the rows and columns in the contingency table, respectively. By taking the minimum of (r – 1, c – 1), the formula scales χ^2 appropriately, avoiding overestimation of the association in situations in which one variable has more categories than the other. The purpose of this scaling factor is 2-fold: (1) It ensures that the Cramer V correlation coefficient, which is the final result, falls within the range of 0-1. This range makes the coefficient interpretable and suitable for comparison across different data sets. (2) It normalizes χ^2

by considering the dimensions of the contingency table (number of rows and columns) and the degrees of freedom. This normalization avoids overestimating the strength of the association in situations in which one variable has more categories than the other.
- Step 3: Calculate the Cramer V correlation coefficient. The final result is a value that ranges from 0 to 1, where 0 indicates no association (variables are independent) and 1 represents a perfect association in the contingency table. This coefficient helps interpret and compare the degree of association across different data sets and contingency tables.

We computed the Cramer V correlation matrix for all pairs of categorical variables in the data set. If the variables had a score of 1, it meant that these variables were representative of the characteristic of the OD. These highly related variables should certainly be represented in the synthetic data for fidelity, which is a statistical measure of similarity. In other words, a Crammer V score of 1 was the threshold and variables scoring 1 were used as the division criteria.

Logical STD in Health Care

National clinical databases differ based on the organization, but clinical data sets are valuable resources [17] that provide insights into improving patient care and organizational efficiency. However, the quality and quantity of clinical data can be limited, especially in cases where data privacy concerns restrict access to real-world data sets. SDG has emerged as a promising solution to this problem, enabling organizations to create new data sets that capture the characteristics of real-world data accurately. However, illogical STDs are frequently generated when simply designed GAN models are used, which induces learning of irregular relationships between the main attribute and its subattributes.

Divide-and-Conquer Approach for Logical STD

As mentioned in the previous section, CS can be a useful approach for generating synthetic data. However, it suffers from the risk of information loss owing to the dependence on condition columns. This is particularly pertinent in cases involving a tabular health care data set, because each of its columns contains significant information [21]. To address these issues, we proposed a DC-based alternative approach.

DC is an easily implementable computing approach [22]. It divides an original problem into several subproblems, analyzes them separately, and then combines the results to obtain the overall solution [23]. DC can be used to generate high-quality synthetic data by dividing the OD set into smaller subsets based on a set of predefined division criteria. This facilitates the specification of complex or multidimensional conditions, while simultaneously reducing the risk of information loss. We followed these steps in the DC approach to generate high-quality synthetic data:

- Step 1: To define the division criteria, we used a methodology involving a class-specific criterion and the Cramer V correlation coefficient. This approach enabled the selection of a different feature subset for each class and consideration of the relationships between different variables to determine the degrees of association.
- Step 2: Based on the defined division criteria, we divided the OD into subsets containing each criterion separately as a specific pattern or relationship. Subsequently, these subsets were used to train GANs on specific patterns and relationships. As a result, the generated STD preserved the patterns and relationships of each subset.
- Step 3 (conquer): The synthetic data corresponding to the different subsets were combined. The generated STD preserved the underlying patterns and relationships within each subset of the OD.

This DC-based approach enabled STD that reflected the underlying patterns and relationships within each subset of the OD accurately.

Generation and Verification of STD

Generative Adversarial Networks

In this study, we used 2 generative models, the CTGAN and CopulaGAN, to generate synthetic data:

- The CTGAN is specifically designed for generating synthetic data from tabular data. The CTGAN exhibits several unique features, including the ability to handle discrete and continuous features, the use of conditional generators, and sampling-specific training to avoid mode collapse and data imbalance problems.
- CopulaGAN uses copulas to model the joint distribution of input features. Copulas are statistical models that describe the dependence structures between random variables, and they have been demonstrated to be effective in modeling complex dependencies between features in real-world data sets.

Prediction Methods

In this study, we validated mortality prediction performance using 4 different classifiers: DT, RF, XGBoost, and LGBM. We used these classifiers to train the ML models and evaluated their performances in predicting mortality in our data set. A sufficiently large training data set was generated in the experiment, and the 4 ML algorithms were used to generate mortality prediction models for patients with non–small cell lung cancer (NSCLC).

- The DT [24] is a commonly used tool. Essentially, a DT is a supervised model that classifies or performs predictions on data sets based on rules in the data. To reach a decision, a DT learns by posing binary questions, which can be represented using a tree structure. The data set is divided hierarchically to contain the greatest amount of information, while branching from the root node. The data are split repeatedly until each segmented region contains a single target value.
- The RF [25] was developed by Leo Breiman and Adele Cutler. It is an extension of the bagging method, which combines the output of multiple DTs to yield a single result. In other words, DTs consider all possible feature splits, while RFs only select a subset of these features. Each tree in an RF ensemble consists of a training set with bootstrap samples. One-third of it is set aside as testing data, known as the out-of-bag (OOB) sample. For a regression task,

individual DTs are averaged, and for a classification task, a majority vote is used to obtain the predicted class. Finally, the OOB sample is used for cross-validation.

- XGBoost [26], a scalable tree-boosting system, is used to solve both classification and regression problems and is a popular algorithm because of its good performance and resource efficiency. XGBoost was developed to handle sparse data. It is an innovative tree learning algorithm that handles instance weights in inexact tree learning, which is a justified weighted quantile sketch procedure. XGBoost enables parallel and distributed computing, which accelerates both learning and model exploration. It exploits out-of-core computation, which enables the construction of an end-to-end system.
- The LGBM [27] is a tree-based learning algorithm with a gradient-boosting framework. In an LGBM, the tree expands vertically compared to other algorithms, in which it expands horizontally. In other words, an LGBM uses a leaf-wise structure, while other algorithms use level-wise structures. An LGBM chooses a leaf with the maximum delta loss to expand, enabling the leaf-wise algorithm to reduce greater loss than its level-wise counterparts.

Experimental Setting

Data Set

Study Population

The Korea Association for Lung Cancer Registry (KALC-R) was developed in cooperation with the Korean Central Cancer Registry and the Lung Cancer Registration Committee. Approximately 10% of NSCLC cases listed in this registry were surveyed in this study. The survey population comprised 13 regional cancer centers and 39 hospitals with numerous registrations [28,29]. Our study used a nonduplicate sample comprising data of 5281 subjects obtained from the KALC-R 2014 and 2015 data sets. Entries with missing and unknown values for weight, height, forced vital capacity (FVC), diffusing capacity of the lungs for carbon monoxide (DLCO), the chemotherapy tumor, extent of spread to lymph nodes, and presence of metastasis (TNM) stage (n=1773, 33.6%), and NSCLC (n=1204, 22.8%) were excluded. This study population (N=2304) was then divided into a development group (n=1616, 70.1%) and a validation group (n=688, 29.9%) via stratified random sampling. The development group used GAN learning for STD and model training for short-term prediction models. The validation group evaluated model performance in terms of ML models in accordance with the quality of prediction. The primary endpoint was defined to be 1 year after the diagnosis of NSCLC for all causes of death.

Moreover, we selected 2 well-known publicly available data sets: the breast cancer data set from the University of California, Irvine (UCI) *Machine Learning Repository* [30] and the diabetes data set [31]. The breast cancer data set comprises real patient data obtained from the Institute of Oncology, Ljubljana, in 1988, aimed at predicting the recurrence of breast cancer. The diabetes data set describes the clinical care at 130 US hospitals and integrated delivery networks from 1999 to 2008. The classification task predicts whether a patient will be readmitted within 30 days.

Comparison of Basic Characteristics

We analyzed the fundamental characteristics of the data sets of patients of NSCLC, breast cancer, and diabetes and further compared the following basic characteristics of different groups for NSCLC survival, breast cancer recurrence, and diabetes readmission in the development data:

- NSCLC: The NSCLC data set exhibited similar distributions across various variables, including age, height, weight, FVC, forced expiratory volume in 1 second (FEV1), DLCO, smoking history (pack-years), gender, Eastern Cooperative Oncology Group (ECOG) performance status, pathological type, epidermal growth factor receptor (EGFR) mutation status, anaplastic lymphoma kinase immunohistochemistry (ALK IHC), anaplastic lymphoma kinase fluorescence in situ hybridization (ALK FISH), cancer stage, curative operations, radiotherapy (RT), chemotherapy, and cause of death. The survival group exhibited lower values for age, height, smoking history (past and current), ECOG performance status, specific cancer types (squamous cell carcinoma, large-cell carcinoma), cancer stage, and palliative chemotherapy compared to the death group. Conversely, the survival group had higher values for weight, FVC, FEV1, DLCO, DLCO percentage, nonsmoking status, adenocarcinoma, positive EGFR mutation, positive ALK IHC, positive ALK FISH, curative operations, RT, and curative chemotherapy.
- Breast cancer: The breast cancer data set also showed comparable distributions for variables, such as age, menopausal status, tumor size, invasive/involved (inv) nodes, node caps, malignancy degree, breast location, breast quadrant, irradiation, and recurrence events. The recurrence group had lower values for age, early menopause (at or before age 40 years), tumor size, inv nodes, node caps, lower malignancy degrees (1 and 2), right breast, breast quadrant, and irradiation compared to the nonrecurrence group. In contrast, the recurrence group had higher values for premenopausal status, malignancy degree (3), and left breast.
- Diabetes: In the diabetes data set, basic characteristics revealed similar distributions for variables, such as hospital stay duration, laboratory procedures, medications, outpatient visits, emergency visits, inpatient stays, diagnoses, race, gender, age, medical specialty, glycated hemoglobin (A1C) results, diabetes medications, and readmission events. The readmitted group displayed lower values for the number of procedures, certain demographics (African American and other races, males), age, medical specialty (except others), A1C result (except none), insulin usage, changes in treatment, and certain diagnoses compared to the nonreadmitted group. Conversely, the readmitted group showed higher values for time spent in the hospital, number of lab procedures, number of medications, number of outpatient visits, number of emergency visits, number of inpatient stays, number of diagnoses, Caucasian race, females, other medical specialties, no A1C result, and diabetes medication (metformin, glipizide, glyburide) usage.

A detailed comparison of the characteristics of different data sets is presented in Multimedia Appendix 1.

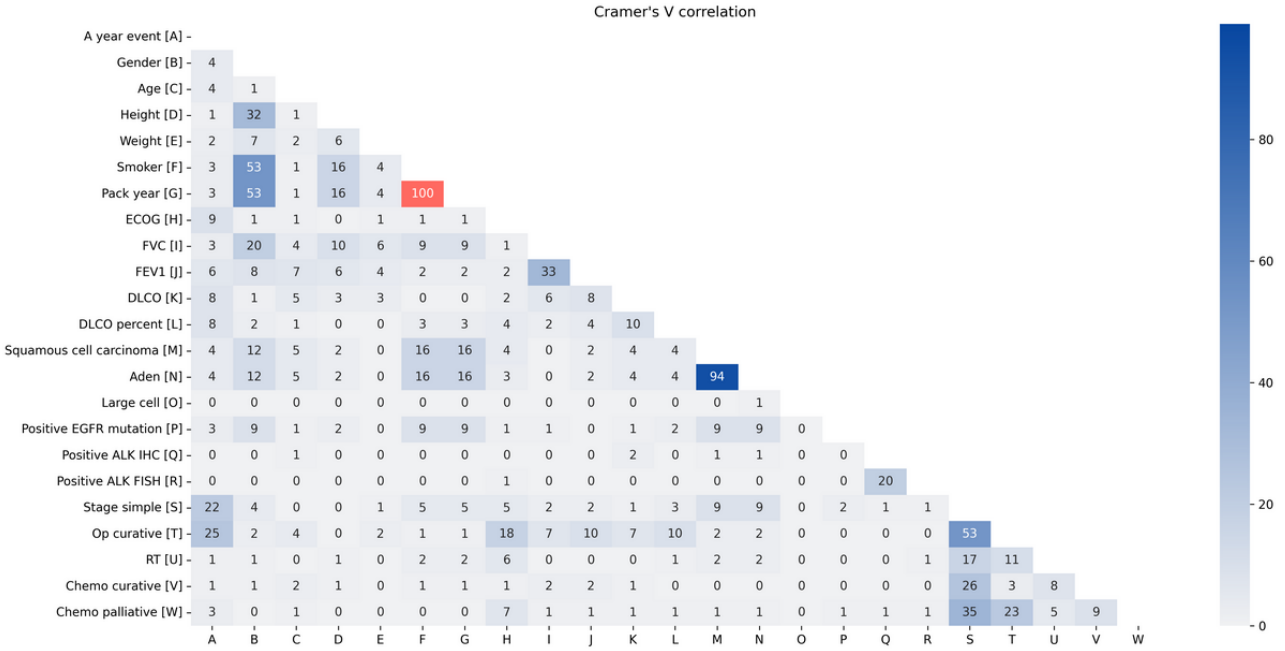
Data Split

Division Criteria Analysis

First, we used class-specific division on the "adverse event" feature, which represents dependent classes, and divided the data into death and survival groups. Next, the Cramer V correlation coefficient was applied after converting all variables into the categorical format. The highest correlation score (V=1, highlighted in red in Figure 2) in NSCLC data was observed between smoking status and pack-years of smoking. This indicates a strong association between these 2 variables, suggesting that individuals who smoke more frequently are more likely to be current or former smokers. Therefore, the "smoker" feature was identified as a key division criterion in our data set. Following the definition, we created a subsample consisting of only those patients in the data set who were smokers and had a pack-year of more than 0. In conclusion, the data were divided into distinct smoker and nonsmoker groups. In other data sets, we did not find a high correlation score, as seen in Multimedia Appendix 2.

By applying the aforementioned division criteria, we obtained 4 small samples from the data set: death-smoker, death-nonsmoker, survival-smoker, and survival-nonsmoker. These samples were used for further performance validation and fidelity tests. Finally, our data set was successfully partitioned for the purposes of our study.

Figure 2. Cramer V correlation coefficient of lung cancer data. Aden: adenocarcinoma; ALK FISH: anaplastic lymphoma kinase fluorescence in situ hybridization; ALK IHC: anaplastic lymphoma kinase immunohistochemistry; chemo curative: curative chemotherapy; chemo palliative; palliative chemotherapy; DLCO: diffusing capacity of the lungs for carbon monoxide; ECOG: Eastern Cooperative Oncology Group; EGFR: epidermal growth factor receptor; FEV1: forced expiratory volume in 1 second; FVC: forced vital capacity; OP curative: curative operations; RT: radiotherapy.



Metrics

Performance Evaluation Metrics

The ability of the synthetic data to achieve good predictive performance in downstream modeling tasks was evaluated using metrics, such as the area under the curve (AUC) and the F_1-score. This is important as the generated synthetic data must be useful for predictive modeling for it to lead to actionable insights.

The AUC is a performance metric that measures the ability of a binary classifier to distinguish between positive and negative classes. It is calculated as the area under the receiver operating characteristic (ROC) [32] curve. ROC curves are graphical representations of the relationship between the false-positive rate (FPR) and the true-positive rate (TPR), plotted along the x and y axes, respectively. The AUC ranges from 0 to 1, where 1 represents perfect classification performance and 0.5 indicates perfectly random performance. The formula for the AUC is given by Equation 2:

The F_1-score is a measure of the balance between precision and recall, where precision is defined as the fraction of true positives among all predicted positives and recall is defined as the fraction of true positives among all actual positives. The F_1-score ranges from 0 to 1, where 1 represents perfect precision and recall and 0 represents the worst-possible scores. The formula for the F1-score is given by Equation 3:

Quality Evaluation Metrics

Shape and pair trend metrics [33] are commonly used to evaluate the fidelity of STD, that is, their similarity to the distribution of real-world data. Shape refers to the overall distributional shape of a data set, including factors such as the degree of skewness or kurtosis. Pair trend, in contrast, refers to the relationship between pairs of features in the data set. Although shape analysis focuses on individual features of a data set, pair

trend analysis provides information about the overall structure and relationships between features. To evaluate the distribution shapes of numerical columns, we used the Kolmogorov-Smirnov statistic (KSS), which is defined as the maximum difference between the cumulative distribution functions (CDFs). CDFs determine the probability that a random observation taken from the population will be less than or equal to a certain value. Conversely, for categorical columns, we used the total variation distance (TVD). The formulas for KSS and TVD scores are given by Equations 4 and 5, respectively, where x represents a single column. Similarly, pair trend metrics were considered to consist of 2 measures, correlation similarity and contingency similarity, for numerical and categorical columns, respectively. Equations 6 and 7 present the formulas for correlation and contingency similarity scores, respectively, where x and y together denote a pair of columns.

By computing separate scores for each column and pair of columns, an individual score was obtained for every column. The final score (value between 0 and 1, with a higher score representing higher quality) was obtained by averaging individual scores. These statistical metrics assessed the similarity or dissimilarity between the distributions of samples in the OD and STD. They provided quantitative measures for evaluating how closely the data sets matched in terms of distribution shapes, relationships between variables, and contingency structures. The final aggregated score represented the overall quality or fidelity of the OD compared to the STD. All 4 measures used for evaluating the fidelity of the OD compared to the STD are summarized in Table 1.

Table 1. Comparison of measures for evaluating fidelity between the OD[a] and the STD[b].

Measure	Purpose	Application	Computation	Interpretation
KSS[c]	Measures the similarity and dissimilarity of distribution shapes between OD and STD	Numerical	Calculates the maximum difference between the CDFs[d] of OD and STD	A higher score indicates greater dissimilarity in distribution shapes, with 0 representing identical distributions.
TVD[e]	Quantifies the difference between probability distributions of categorical data in OD and STD	Categorical	Measures the "closeness" between probability mass functions of OD and STD distributions	A score of 0 implies identical distributions, while higher scores indicate increasing dissimilarity.
Correlation	Evaluates the similarity or dissimilarity of relationships between pairs of numerical variables between OD and STD	Numerical	Measures the absolute squared difference between correlation coefficients of OD and STD pairs	A score of 0 indicates perfect similarity in relationships, while higher scores suggest weaker similarity or even dissimilarity.
Contingency	Assesses the similarity of relationships between pairs of categorical variables in OD and STD	Categorical	Calculates the sum of absolute differences between corresponding cells in contingency tables (cross-tabulations) of OD and STD	A score of 0 signifies perfect similarity in contingency structures, while higher scores indicate less similarity.

[a]OD: original data.

[b]STD: synthetic tabular data.

[c]KSS: Kolmogorov-Smirnov statistic.

[d]CDF: cumulative distribution function.

[e]TVD: total variation distance.

Results

Generation and Validation of STD

To generate logical STD, we trained the CTGAN and CopulaGAN using existing CS filtering. Next, we used the proposed DC-based method before training the CTGAN and CopulaGAN without CS filtering. The volume of the generated data set was set to 5000. Moreover, we generated 2 types of STDs, a balanced data set with equal class distributions between samples in a 50:50 ratio and an imbalanced data set with a 1:100 class distribution ratio between samples (ie, each dependent variable occurred 100 times less frequently than its counterpart).

To verify the superiority of the proposed DC-based method in the generation of logical STD, we evaluated each STD item using 4 different ML models (DT, RF, XGBoost, and LGBM). Table 2 presents the validation results of the DT classifier. The AUC and F_1-score values of the NSCLC, breast cancer, and diabetes OD were 66.06% and 66.11%, 61.14% and 49.64%, and 65.58% and 47.82%, respectively. The highest AUC of 74.87% was achieved by generating synthetic data using the DC strategy with the CopulaGAN, while the highest F_1-score of 71.99% was achieved using the DC strategy with the CTGAN for NSCLC data. The highest AUC of 73.31% was achieved by generating synthetic data using the DC strategy with the CTGAN, while the highest F_1-score of 68.92% was achieved using the DC strategy with the CopulaGAN for breast cancer data. The highest AUC of 61.57% was achieved by generating synthetic data using the DC strategy with the CTGAN, while the highest F_1-score of 53.8% was achieved using the DC strategy with the CopulaGAN for diabetes data.

The validation results obtained using the RF classifier are presented in Table 3. The AUC and F_1-score values of the NSCLC, breast cancer, and diabetes OD were 84.81% and 72.74%, 69.37% and 60.01%, and 62.13% and 47.73%, respectively. The highest AUC and F_1-score of 85.61% and 75.09%, respectively, were achieved by generating synthetic

data using the DC strategy with the CTGAN for NSCLC data. The highest AUC and F_1-score of 78.05% and 71.03%, respectively, were achieved by generating synthetic data using the DC strategy with the CTGAN for breast cancer data. The highest AUC and F_1-score of 59.98% and 53.47%, respectively, were achieved by generating synthetic data using the DC strategy with the CTGAN for diabetes data.

Table 4 presents the validation results obtained using the XGBoost classifier. The AUC and F_1-score values of the NSCLC, breast cancer, and diabetes OD were 83.07% and 71.14%, 71.21% and 62.89%, and 67.02% and 48.91%, respectively. The highest AUC and F_1-score of 85.20% and 74.78%, respectively, were achieved by generating synthetic data using the DC strategy with the CTGAN for NSCLC data. The highest AUC and F_1-score of 77.86% and 70.58%, respectively, were achieved by generating synthetic data using the DC strategy with the CTGAN for breast cancer data. The highest AUC and F_1-score of 60.18% and 53.93%, respectively, were achieved by generating synthetic data using the DC strategy with the CTGAN for diabetes data.

Finally, Table 5 presents the validation results obtained using the LGBM classifier. The AUC and F_1-score values of the NSCLC, breast cancer, and diabetes OD were 84.09% and 71.30%, 75.84% and 62.07%, and 67.88% and 47.89%, respectively. The highest AUC and F_1-score of 85.14% and 74.40%, respectively, were achieved by generating synthetic data using the DC strategy with the CTGAN for NSCLC data. The highest AUC and F_1-score of 77.86% and 70.58%, respectively, were achieved by generating synthetic data using the DC strategy with the CTGAN for breast cancer data. The highest AUC and F_1-score of 60.18% and 53.93%, respectively, were achieved by generating synthetic data using the DC strategy with the CTGAN for diabetes data.

In general, the results demonstrate that STD generated using the DC approach had the best quality in terms of the AUC and F_1-score. Moreover, higher performance was observed when STD were generated solely using the DC approach compared to STD obtained using the original training data. Moreover, balanced data sets consistently exhibited better performance than imbalanced ones.

In addition, we assessed the quality of the generated STD by evaluating their fidelity with respect to shape and pair trend metrics. The results are presented in Tables 6-8. The DC strategy with the CTGAN achieved the highest mean shape score of 90.49 (SD 0.07), 91.71 (SD 0.12), and 98.60 (SD 0.13), the highest mean pair trend score of 83.92 (SD 0.10), 82.72 (SD 0.13), and 96.70 (SD 0.26), and the highest mean overall score of 87.20 (SD 0.08), 87.21 (SD 0.09), and 97.65 (SD 0.27) on the NSCLC, breast cancer, and diabetes data sets, respectively. These findings suggest that the DC strategy with the CTGAN could be a promising approach for generating synthetic data with high fidelity. Moreover, we carried out a number of visualization experiments comparing the OD and the STD, as shown in Multimedia Appendix 3.

Table 2. Validation results obtained using the DT[a] classifier: mean (SD) values of 5 experiments.

Data type, GAN[b], and approach	NSCLC[c]		Breast cancer		Diabetes	
	AUC[d]	F_1-score	AUC	F_1-score	AUC	F_1-score
OD[e]	66.06 (1.31)	66.11 (1.30)	61.14 (1.66)	49.64 (1.06)	65.58 (0.13)	47.82 (0.04)
Balanced STD[f]						
CTGAN[g], no condition	63.46 (1.91)	61.87 (1.83)	59.10 (2.17)	62.20 (1.93)	57.19 (0.25)	49.16 (0.10)
CTGAN, CS[h]	59.95 (1.26)	58.65 (2.44)	64.65 (1.66)	56.67 (1.31)	58.93 (0.18)	47.89 (0.09)
CTGAN, DC[i]	74.50 (1.23)	71.99 (0.77)[j]	73.31 (1.11)[j]	67.10 (1.83)	61.57 (0.09)[j]	50.45 (0.05)
CopulaGAN[k], no condition	66.11 (1.32)	65.16 (1.27)	59.82 (1.69)	52.32 (2.11)	58.93 (0.24	47.13 (0.04)
CopulaGAN, CS	63.87 (2.02)	63.07 (1.87)	67.96 (2.15)	62.20 (1.93)	60.08 (0.17)	49.77 (0.04)
CopulaGAN, DC	74.87 (0.77)[j]	70.54 (0.76)	70.66 (0.85)	68.92 (1.78)[j]	60.61 (0.06)	53.80 (0.08)[j]
Imbalanced STD						
CTGAN, no condition	50.93 (0.89)	44.99 (1.97)	52.48 (0.09	38.75 (0.51	57.60 (0.14	47.09 (0.01
CTGAN, CS	50.64 (0.80)	44.80 (1.48)	52.48 (0.09)	38.75 (0.51)	58.34 (0.24)	47.09 (0.01)
CTGAN, DC	57.99 (2.06)	57.78 (2.81)	56.00 (0.31)	38.75 (0.51)	60.52 (0.05)	50.36 (0.10)
CopulaGAN, no condition	52.32 (1.05)	48.38 (1.92)	53.81 (0.63)	38.75 (0.51)	57.10 (0.21)	47.09 (0.01)
CopulaGAN, CS	52.06 (0.91)	47.62 (1.70)	55.50 (1.01)	38.75 (0.51)	56.21 (0.24)	47.09 (0.01)
CopulaGAN, DC	55.86 (3.10)	54.12 (4.60)	57.95 (0.76)	38.75 (0.51)	59.38 (0.15)	50.22 (0.08)

[a]DT: decision tree.

[b]GAN: generative adversarial network.

[c]NSCLC: non–small cell lung cancer.

[d]AUC: area under the curve.

[e]OD: original data.

[f]STD: synthetic tabular data.

[g]CTGAN: conditional tabular generative adversarial network.

[h]CS: conditional sampling.

[i]DC: divide and conquer.

[j]The best results.

[k]CopulaGAN: copula generative adversarial network.

Table 3. Validation results obtained using the RF[a] classifier: mean (SD) values of 5 experiments.

Data type, GAN[b], and approach	NSCLC[c]		Breast cancer		Diabetes	
	AUC[d]	F_1-score	AUC	F_1-score	AUC	F_1-score
OD[e]	84.81 (0.23)	72.74 (0.30)	69.37 (1.22)	60.01 (1.09)	62.13 (0.48)	47.73 (0.16)
Balanced STD[f]						
CTGAN[g], no condition	79.07 (1.12)	67.70 (1.54)	67.73 (3.24)	65.96 (4.37)	56.93 (0.31)	46.59 (0.16)
CTGAN, CS[h]	79.01 (1.20)	68.47 (1.39)	54.88 (3.00)	52.59 (2.74)	57.23 (0.33)	46.38 (0.19)
CTGAN, DC[i]	85.61 (0.29)[j]	75.09 (0.58)[j]	78.05 (1.59)[j]	71.03 (2.11)[j]	59.98 (0.24)[j]	53.47 (0.13)[j]
CopulaGAN[k], no condition	78.29 (0.74)	67.28 (1.67)	59.14 (2.32)	57.16 (3.46)	56.92 (0.26)	44.17 (0.12)
CopulaGAN, CS	78.16 (1.72)	68.82 (1.82)	73.48 (4.73)	64.48 (8.02)	58.55 (0.24)	44.80 (0.13)
CopulaGAN, DC	83.91 (0.35)	72.97 (0.67)	77.82 (1.83)	66.61 (4.66)	58.27 (0.31)	52.86 (0.26)
Imbalanced STD						
CTGAN, no condition	67.20 (2.74)	41.94 (0.00)	53.29 (5.42)	38.75 (0.51)	53.18 (0.26)	47.09 (0.01)
CTGAN, CS	68.03 (1.26)	41.94 (0.00)	54.48 (2.75)	38.75 (0.51)	54.41 (0.24)	47.09 (0.01)
CTGAN, DC	77.98 (1.12)	47.05 (1.68)	59.81 (1.47)	39.84 (2.60)	56.44 (0.55)	49.42 (0.07)
CopulaGAN, no condition	67.70 (2.00)	41.96 (0.11)	54.59 (1.20)	38.75 (0.51)	52.77 (0.35)	47.09 (0.01)
CopulaGAN, CS	67.70 (1.84)	41.96 (0.11)	55.83 (2.32)	38.75 (0.51)	53.19 (0.76)	47.09 (0.01)
CopulaGAN, DC	78.73 (1.57)	44.63 (0.88)	58.61 (1.96)	38.75 (0.51)	55.20 (0.23)	48.74 (0.68)

[a]RF: random forest.

[b]GAN: generative adversarial network.

[c]NSCLC: non–small cell lung cancer.

[d]AUC: area under the curve.

[e]OD: original data.

[f]STD: synthetic tabular data.

[g]CTGAN: conditional tabular generative adversarial network.

[h]CS: conditional sampling.

[i]DC: divide and conquer.

[j]The best results.

[k]CopulaGAN: copula generative adversarial network.

Table 4. Validation results obtained using the XGBoost[a] classifier: mean (SD) values of 5 experiments.

Data type, GAN[b], and approach	NSCLC[c]		Breast cancer		Diabetes	
	AUC[c]	F_1-score	AUC	F_1-score	AUC	F_1-score
OD[d]	83.07 (0.37)	71.14 (1.09)	71.21 (0.46)	62.89 (2.59)	67.02 (0.13)	48.91 (0.18)
Balanced STD[e]						
CTGAN[g], no condition	76.50 (1.15)	66.48 (1.75)	68.77 (2.49)	68.38 (4.49)	57.95 (0.30)	47.85 (0.22)
CTGAN, CS[h]	74.71 (1.59)	64.86 (1.75)	54.15 (2.49)	48.44 (3.25)	58.31 (0.20)	47.54 (0.15)
CTGAN, DC[i]	85.20 (0.82)[j]	74.78 (0.77)[j]	77.86 (2.27)[j]	70.58 (3.36)[j]	60.18 (0.20)[j]	53.93 (0.29)[j]
CopulaGAN[k], no condition	77.18 (0.98)	67.56 (1.63)	56.47 (2.12)	54.75 (2.39)	59.25 (0.21)	46.51 (0.12)
CopulaGAN, CS	76.42 (0.93)	67.82 (1.22)	68.32 (2.37)	60.65 (5.21)	58.98 (0.29)	46.55 (0.29)
CopulaGAN, DC	83.58 (0.65)	72.92 (0.66)	77.69 (1.91)	64.96 (2.87)	58.84 (0.35)	53.10 (0.34)
Imbalanced STD						
CTGAN, no condition	72.18 (4.12)	42.12 (0.42)	59.76 (1.72)	38.75 (0.51)	54.31 (0.30)	47.09 (0.01)
CTGAN, CS	70.94 (2.93)	42.07 (0.32)	61.65 (1.49)	38.75 (0.51)	56.93 (0.37)	47.09 (0.01)
CTGAN, DC	83.20 (0.42)	62.13 (2.43)	70.06 (1.07)	38.75 (0.51)	59.35 (0.39)	49.15 (0.29)
CopulaGAN, no condition	72.40 (3.23)	42.59 (0.45)	65.77 (2.03)	38.75 (0.51)	55.86 (0.39)	47.10 (0.02)
CopulaGAN, CS	73.21 (1.81)	42.75 (0.46)	57.42 (1.04)	38.75 (0.51)	54.63 (0.23)	47.09 (0.01)
CopulaGAN, DC	82.60 (1.37)	59.22 (1.78)	68.19 (2.85)	38.75 (0.51)	57.95 (0.49)	48.26 (0.17)

[a]XGBoost: Extreme Gradient Boosting.

[b]GAN: generative adversarial network.

[c]NSCLC: non–small cell lung cancer.

[d]AUC: area under the curve.

[e]OD: original data.

[f]STD: synthetic tabular data.

[g]CTGAN: conditional tabular generative adversarial network.

[h]CS: conditional sampling.

[i]DC: divide and conquer.

[j]The best results.

[k]CopulaGAN: copula generative adversarial network.

Table 5. Validation results obtained using the LGBM[a] classifier: mean (SD) values of 5 experiments.

Data type, GAN[b], and approach	NSCLC[c]		Breast cancer		Diabetes	
	AUC[d]	F_1-score	AUC	F_1-score	AUC	F_1-score
OD[e]	84.09 (0.03)	71.30 (0.72)	75.84 (1.80)	62.07 (3.05)	67.88 (0.12)	47.89 (0.16)
Balanced STD[f]						
CTGAN[g], no condition	77.31 (1.21)	66.90 (1.65)	66.52 (1.60)	65.32 (2.87)	58.42 (0.26)	48.38 (0.16)
CTGAN, CS[h]	77.43 (1.78)	67.25 (2.09)	57.94 (3.23)	60.42 (3.45)	58.75 (0.17)	48.31 (0.07)
CTGAN, DC[i]	85.14 (0.70)[j]	74.40 (0.78)[j]	78.16 (1.52)[j]	71.75 (1.79)[j]	61.75 (0.13)[j]	54.09 (0.19)[j]
CopulaGAN[k], no condition	77.61 (0.86)	68.50 (1.21)	59.38 (1.15)	56.21 (3.63)	60.36 (0.27)	46.68 (0.12)
CopulaGAN, CS	77.62 (1.85)	68.58 (1.14)	70.02 (2.17)	65.51 (2.97)	61.12 (0.23)	46.75 (0.12)
CopulaGAN, DC	83.57 (0.55)	72.84 (0.59)	75.31 (2.45)	68.13 (1.72)	60.03 (0.23)	53.63 (0.16)
Imbalanced STD						
CTGAN, no condition	71.84 (3.03)	42.03 (0.20)	59.81 (0.66)	38.75 (0.51)	55.79 (0.25)	47.09 (0.01)
CTGAN, CS	71.51 (2.65)	41.96 (0.11)	64.99 (1.48)	38.75 (0.51)	58.47 (0.24)	47.10 (0.01)
CTGAN, DC	83.60 (1.00)	61.55 (2.68)	70.54 (0.56)	43.00 (0.68)	60.80 (0.21)	49.71 (0.06)
CopulaGAN, no condition	73.83 (3.15)	42.27 (0.40)	63.92 (1.67)	38.75 (0.51)	56.45 (0.34)	47.10 (0.03)
CopulaGAN, CS	74.21 (2.37)	42.42 (0.53)	60.28 (0.73)	38.75 (0.51)	56.31 (0.29)	47.11 (0.03)
CopulaGAN, DC	81.92 (1.10)	57.91 (2.42)	72.84 (1.40)	40.83 (2.71)	58.34 (0.35)	48.53 (0.11)

[a]LGBM: light gradient-boosting machine.

[b]GAN: generative adversarial network.

[c]NSCLC: non–small cell lung cancer.

[d]AUC: area under the curve.

[e]OD: original data.

[f]STD: synthetic tabular data.

[g]CTGAN: conditional tabular generative adversarial network.

[h]CS: conditional sampling.

[i]DC: divide and conquer.

[j]The best results.

[k]CopulaGAN: copula generative adversarial network.

Table 6. Summary of quality tests for the NSCLC[a] data set: mean (SD) values of 5 experiments.

Approach and GAN[b]	Shape	Pair trend	Overall
CS[c]			
CTGAN[d]	88.42 (0.12)	83.47 (0.10)	85.95 (0.11)
CopulaGAN[e]	89.98 (0.06)	79.77 (3.84)	84.87 (1.93)
DC[f]			
CTGAN	90.49 (0.07)[g]	83.92 (0.10)[g]	87.20 (0.08)[g]
CopulaGAN	89.72 (0.07)	82.48 (0.12)	86.10 (0.09)

[a]NSCLC: non–small cell lung cancer.
[b]GAN: generative adversarial network.
[c]CS: conditional sampling.
[d]CTGAN: conditional tabular generative adversarial network.
[e]CopulaGAN: copula generative adversarial network.
[f]DC: divide and conquer.
[g]The best results.

Table 7. Summary of quality tests for the breast cancer data set: mean (SD) values of 5 experiments.

Approach and GAN[a]	Shape	Pair trend	Overall
CS[b]			
CTGAN[c]	90.75 (0.14)	80.97 (0.17)	85.86 (0.15)
CopulaGAN[d]	89.25 (0.18)	80.68 (0.16)	84.97 (0.21)
DC[e]			
CTGAN	91.71 (0.12)[f]	82.72 (0.13)[f]	87.21 (0.09)[f]
CopulaGAN	91.18 (0.11)	81.24 (0.14)	86.21 (0.14)

[a]GAN: generative adversarial network.
[b]CS: conditional sampling.
[c]CTGAN: conditional tabular generative adversarial network.
[d]CopulaGAN: copula generative adversarial network.
[e]DC: divide and conquer.
[f]The best results.

Table 8. Summary of quality tests for the diabetes data set: mean (SD) values of 5 experiments.

Approach and GAN[a]	Shape	Pair trend	Overall
CS[b]			
CTGAN[c]	97.55 (0.23)	95.27 (0.32)	96.41 (0.26)
CopulaGAN[d]	97.22 (0.24)	94.27 (0.32)	95.74 (0.36)
DC[e]			
CTGAN	98.60 (0.31)[f]	96.70 (0.26)[f]	97.65 (0.27)[f]
CopulaGAN	97.90 (0.27)	95.74 (0.23)	96.82 (0.31)

[a]GAN: generative adversarial network.

[b]CS: conditional sampling.

[c]CTGAN: conditional tabular generative adversarial network.

[d]CopulaGAN: copula generative adversarial network.

[e]DC: divide and conquer.

[f]The best results.

Discussion

Principal Findings

Preserving data with logical relationships while generating STD using GANs has not been sufficiently researched. Some GANs, such as the CTGAN and CopulaGAN, use CS filtering to determine the exclusion of record data based on predefined condition columns after generating STD. However, this is highly dependent on condition columns, which may lead to meaningful information in the excluded records being ignored. To resolve this problem, we proposed a DC-based approach in this paper, as shown in Multimedia Appendix 4.

The proposed DC-based approach was verified to produce STD involving logical relationships between columns. As the division strategy, we used class-specific and the Cramer V criteria sequentially. First, we used a class-specific criterion to classify dependent classes between survival and death groups. Subsequently, we measured the relative degrees of association among pairs of variables based on the Cramer V correlation coefficient in order to identify strong evidence for meaningful correlations between columns. In terms of a high Cramer V correlation coefficient (=1), smoker and nonsmoker groups were selected as division criteria. Using this, the OD was divided into smaller data sets comprising hierarchical group data that considered class-specific aspects of learning. Further, the division criteria of the DC strategy avoided the problem of ignoring some records owing to overreliance on condition columns.

To compare the logical STD generation approaches, we trained the CTGAN and CopulaGAN with CS filtering and compared their performances with those of ML models trained using a DC approach without CS filtering. The results demonstrated that the epochs hyperparameter was sensitive, with a significant impact on the quality of synthetic data generated using the CTGAN and CopulaGAN. Specifically, the results depended considerably on the value of the epochs hyperparameter, ranging from 100 to 500. We used a grid search algorithm to identify an optimal value for the epochs hyperparameter, as shown in Multimedia Appendix 5. Regularization hyperparameters, such as grid search, are essential to the generalization of ML models [34]. They work well with low-dimensional hyperparameter spaces and ample computational resources [35]. A grid search involves testing a range of hyperparameter values and evaluating the performance of the model corresponding to each value. In our case, we tested epoch values of 100, 200, 300, 400, and 500 and evaluated the resulting synthetic data using a variety of metrics, including distributional similarity, feature correlation, and downstream performance in predictive models. Our findings highlight the importance of carefully selecting hyperparameters during GAN training to generate synthetic data from clinical data sets. The sensitivity of the epochs hyperparameter underscores the necessity of systematic approaches, such as grid search, to identify optimal values.

Generally, ML training on imbalanced data sets leads to failure to properly learn the distributive characteristics of the data and, consequently, unfavorable accuracies across the data classes [36]. We generated balanced and imbalanced STD by regulating the volumes of the dependence variables for comparison. These data were used to develop ML models (DT, RF, XGBoost, and LGBM), and their AUC and F_1-score were measured on the verification data set. The hyperparameter of each model was tuned via a grid search for the number of epochs. All balanced synthetic data exhibited higher performance on the prediction models (DT, RF, XGBoost, and LGBM) compared to imbalanced synthetic data. Therefore, we recommend that the volume of balanced dependence variables be considered during SDG using GANs.

Finally, the DC-based approach was observed to exhibit several potential advantages over CS. First, deconstruction of the division criteria into simpler subrules enables the specification of complex or multidimensional conditions. Second, training the GAN on each subrule independently reduces the risk of information loss by CS, as the GAN can focus on generating synthetic data that accurately reflect the distribution of the data for each subrule. Finally, combining the results of the subrules enables the generation of synthetic data that satisfy all the

original logical rules, without requiring complex and potentially overspecified conditions.

Thus, the main contribution of this paper is to demonstrate the viability of the proposed STD generation method to serve as a revolutionary new alternative to existing counterparts in the development of ML-based prediction models.

Limitations

Our study is limited in terms of the low dimensionality and count of data collected from a single country. In practical health care, low-dimensional and sparse data are often derived from strict data preprocessing, a detailed design for the target population, or exact primary endpoints. In this paper, data containing essential variables were collected from 13 regional cancer centers and 39 hospitals via sampling. However, patients with NSCLC from only a single country were considered, potentially introducing racial bias. We intend to overcome this limitation in future works by applying the proposed framework to data collected from other countries.

The DC-based STD learning strategy may be difficult to apply in the case of sparse data and multiple division criteria. Indiscriminate use of the strategy, even in the presence of a large amount of data, can be problematic because the use of multiple division strategies induces a lack of learning data, which motivates the generation of inappropriate synthetic data. Therefore, it is important to establish appropriate criteria for the division strategy (eg, the class-specific and Cramer V criteria proposed in our study). We recommend that the class-specific criterion be used as an essential strategy in the first division criteria. The Cramer V criterion should be used to calculate correlations between variables, enabling sufficient discussion about the group of candidates for division and helping decide the need for division.

One potential challenge with the DC approach is that the subrules and the combinations of results require careful consideration. If the subrules are not well defined or the combinations of results are not appropriate, the resulting synthetic data may not accurately reflect the characteristics of real-world data. Additionally, if data with logical relationships are highly interdependent, it may be challenging to break them down into independent subrules. Despite these potential challenges, the DC approach exhibited great promise in generating synthetic data from data with logical relationships on clinical data sets.

Conclusion

Our study demonstrated problems of CS-based STD generation techniques and the feasibility of DC-based STD generation to address those problems. Further, the effectiveness of the generated STD to enable the application of ML models was verified, revealing that they improve prediction performance.

Acknowledgments

This study was supported by a grant (no: 2310440-1) offered by the National Cancer Center of Korea, Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (no: NRF-2022R1F1A1075041), and the Korea Health Technology R&D Project through the Korea Health Industry Development Institute, funded by the Ministry of Health & Welfare, Republic of Korea (no: HI21C0974).

Data Availability

Anyone can use the original data after registering as a member on the Korea Central Cancer Registry (KCCR) portal [37] and passing through the data application and review. Users need to fill out an application form, including a research proposal describing how they will use the data and that the data access request will be accessed by the KCCR and the National Statistics Office. All synthetic data can be shared for research purposes by contacting the authors.

All code for data generation and validation associated with the current submission is available in a GitHub repository [38].

Authors' Contributions

Conceptualization was managed by HYJK, MSK, and KSR; methodology, HYJK, EB, MSK, and KSR; validation, HYJK, EB, MSK, KSC, and KSR; investigation, HYJK, EB, MSK, and KSR; data curation, HYJK, DWC, and KSR; and writing—original draft preparation HYJK, EB, MSK, and KSR. All authors have assisted in the drafting and editing of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Data set characteristics.
[DOCX File , 28 KB - medinform_v11i1e47859_app1.docx]

Multimedia Appendix 2
Cramer V correlation.
[DOCX File , 451 KB - medinform_v11i1e47859_app2.docx]

Multimedia Appendix 3
Distribution plots.
[DOCX File , 5768 KB - medinform_v11i1e47859_app3.docx]

Multimedia Appendix 4
As-is and to-be.
[DOCX File , 145 KB - medinform_v11i1e47859_app4.docx]

Multimedia Appendix 5
Effect of epoch on validation results.
[DOCX File , 12013 KB - medinform_v11i1e47859_app5.docx]

References

1. Ghassemi M, Naumann T, Schulam P, Beam AL, Chen IY, Ranganath R. A review of challenges and opportunities in machine learning for health. AMIA Jt Summits Transl Sci Proc 2020;2020:191-200 [FREE Full text] [Medline: 32477638]
2. Henry KE, Hager DN, Pronovost PJ, Saria S. A targeted real-time early warning score (TREWScore) for septic shock. Sci Transl Med 2015 Aug 05;7(299):299ra122. [doi: 10.1126/scitranslmed.aab3719] [Medline: 26246167]
3. Zhang A, Xing L, Zou J, Wu JC. Shifting machine learning for healthcare from development to deployment and from models to data. Nat Biomed Eng 2022 Jul 04;6(12):1330-1345. [doi: 10.1038/s41551-022-00898-y] [Medline: 35788685]
4. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial networks. Commun ACM 2020 Oct 22;63(11):139-144. [doi: 10.1145/3422622]
5. Piacentino E, Guarner A, Angulo C. Generating synthetic ECGs using gans for anonymizing healthcare data. Electronics 2021 Feb 05;10(4):389. [doi: 10.3390/electronics10040389]
6. Arora A, Arora A. Generative adversarial networks and synthetic patient data: current challenges and future perspectives. Future Healthc J 2022 Jul;9(2):190-193 [FREE Full text] [doi: 10.7861/fhj.2022-0013] [Medline: 35928184]
7. Gonzales A, Guruswamy G, Smith SR. Synthetic data in health care: a narrative review. PLOS Digit Health 2023 Jan;2(1):e0000082 [FREE Full text] [doi: 10.1371/journal.pdig.0000082] [Medline: 36812604]
8. Azizi Z, Zheng C, Mosquera L, Pilote L, El Emam K, GOING-FWD Collaborators. Can synthetic data be a proxy for real clinical trial data? A validation study. BMJ Open 2021 Apr 16;11(4):e043497 [FREE Full text] [doi: 10.1136/bmjopen-2020-043497] [Medline: 33863713]
9. Lei X, Kalyan V. Synthesizing tabular data using generative adversarial networks. arXiv Preprint posted online 27 November, 2018. [doi: 10.48550/arXiv.1811.11264]
10. Choi E, Biswal S, Malin B, Duke J, Stewart WF, Sun J. Generating multi-label discrete patient records using generative adversarial networks. 2017 Presented at: 2017 MHLC: 2nd Machine Learning for Healthcare Conference; August 18-19, 2017; Boston, MA p. 286-305.
11. Krenmayr L, Frank R, Drobig C, Braungart M, Seidel J, Schaudt D, et al. GANerAid: realistic synthetic patient data for clinical trials. Inform Med Unlocked 2022;35:101118. [doi: 10.1016/j.imu.2022.101118]
12. Lan L, You L, Zhang Z, Fan Z, Zhao W, Zeng N, et al. Generative adversarial networks and its applications in biomedical informatics. Front Public Health 2020 May 12;8:164 [FREE Full text] [doi: 10.3389/fpubh.2020.00164] [Medline: 32478029]
13. Beaulieu-Jones BK, Wu ZS, Williams C, Lee R, Bhavnani SP, Byrd JB, et al. Privacy-preserving generative deep neural networks support clinical data sharing. Circ: Cardiovasc Qual Outcomes 2019 Jul;12(7):e005122. [doi: 10.1161/circoutcomes.118.005122]
14. Baowaly M, Lin C, Liu C, Chen K. Synthesizing electronic health records using improved generative adversarial networks. J Am Med Inform Assoc 2019 Mar 01;26(3):228-241 [FREE Full text] [doi: 10.1093/jamia/ocy142] [Medline: 30535151]
15. Izonin, Ivan. Experimental evaluation of the effectiveness of ann-based numerical data augmentation methods for diagnostics tasks. 2021 Presented at: IDDM-2021: 4th International Conference on Informatics & Data-Driven Medicine; November 19-21, 2021; Valencia, Spain.
16. Wang Z, Myles P, Tucker A. Generating and evaluating synthetic UK primary care data: preserving data utility patient privacy. 2019 Presented at: IEEE CBMS2019: 32nd IEEE International Symposium on Computer-Based Medical Systems (CBMS); June 5-7, 2019; Córdoba, Spain. [doi: 10.1109/cbms.2019.00036]
17. Hammad R, Barhoush M, Abed-Alguni BH. A semantic-based approach for managing healthcare big data: a survey. J Healthc Eng 2020 Nov 22;2020:8865808-8865812 [FREE Full text] [doi: 10.1155/2020/8865808] [Medline: 33489061]
18. Xu L, Skoularidou M, Cuesta-Infante A, Veeramachaneni K. Modeling tabular data using conditional GAN. 2019 Presented at: NeurIPS 2019: 33rd Conference on Neural Information Processing Systems; December 8-14, 2019; Vancouver, Canada.
19. CopulaGAN model. SDV. URL: https://sdv.dev/SDV/user_guides/single_table/copulagan.html [accessed 2023-11-01]
20. Pineda-Bautista BB, Carrasco-Ochoa J, Martı´nez-Trinidad JF. General framework for class-specific feature selection. Expert Syst Appl 2011 Aug;38(8):10018-10024. [doi: 10.1016/j.eswa.2011.02.016]

21. Chong KM. Privacy-preserving healthcare informatics: a review. ITM Web Conf 2021 Jan 26;36:04005. [doi: 10.1051/itmconf/20213604005]
22. Chen X, Cheng JQ, Xie MG. Divide-and-conquer methods for big data analysis. arXiv Preprint posted online 22 February 2021. [doi: 10.48550/arXiv.2102.10771]
23. Chen X, Liu W, Zhang Y. Quantile regression under memory constraint. Ann Stat 2019 Dec 1;47(6):3244-3273. [doi: 10.1214/18-AOS1777]
24. Navada A, Ansari A, Patil S, Sonkamble B. Overview of use of decision tree algorithms in machine learning. 2011 Presented at: ICSGRC 2011: 2011 IEEE Control and System Graduate Research Colloquium; June 27-28, 2011; Shah Alam, Malaysia p. 37-42. [doi: 10.1109/icsgrc.2011.5991826]
25. Breiman L. Random forests. Mach Learn 2001 Oct;45:4-32. [doi: 10.1023/A:1010933404324]
26. Chen T, Guestrin C. Xgboost: a scalable tree boosting system. 2016 Presented at: KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 13-17, 2016; San Francisco, CA p. 785-794. [doi: 10.1145/2939672.2939785]
27. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: a highly efficient gradient boosting decision tree. 2017 Presented at: NIPS 2017: 31st Conference on Neural Information Processing Systems; December 4-9, 2017; Long Beach, CA.
28. Choi C, Kim H, Jung C, Cho D, Jeon J, Lee J, et al. Report of the Korean Association of Lung Cancer Registry (KALC-R), 2014. Cancer Res Treat 2019 Oct;51(4):1400-1410 [FREE Full text] [doi: 10.4143/crt.2018.704] [Medline: 30913875]
29. Jeon DS, Kim S, Kim T, Kim H, Kim HK, Moon MH, Korean Association for Lung Cancer, Korea Central Cancer Registry. Five-year overall survival and prognostic factors in patients with lung cancer: results from the Korean Association of Lung Cancer Registry (KALC-R) 2015. Cancer Res Treat 2023 Jan;55(1):103-111 [FREE Full text] [doi: 10.4143/crt.2022.264] [Medline: 35790197]
30. Dua D, Graff C. UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science; 2017.
31. Strack B, DeShazo J, Gennings C, Olmo J, Ventura S, Cios K, et al. Impact of HbA1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. Biomed Res Int 2014;2014:781670 [FREE Full text] [doi: 10.1155/2014/781670] [Medline: 24804245]
32. Fawcett T. An introduction to ROC analysis. Pattern Recogn Lett 2006 Jun;27(8):861-874 [FREE Full text] [doi: 10.1016/j.patrec.2005.10.010] [Medline: 24804245]
33. Synthetic data metrics. Version 0.9.3. DataCebo. 2023 Apr. URL: https://docs.sdv.dev/sdmetrics/ [accessed 2023-10-06]
34. Feurer M, Hutter F. Hyperparameter optimization. In: Hutter F, Kotthoff L, Vanschoren J, editors. Automated Machine Learning. The Springer Series on Challenges in Machine Learning. Cham: Springer; 2019:3-33.
35. Bergstra J, Bengio Y. Random search for hyper-parameter optimization. J Mach Learn Res 2012 Feb 1;13(2):281-305.
36. He H, Garcia EA. Learning from imbalanced sata. IEEE Trans Knowl Data Eng 2009 Sep;21(9):1263-1284. [doi: 10.1109/tkde.2008.239]
37. Home page. Korea Central Cancer Registry. URL: https://kccrsurvey.cancer.go.kr/index.do [accessed 2023-11-01]
38. Kwang SR. Wally-AI/DC. GitHub. URL: https://github.com/KwangSun-Ryu/Wally-AI/tree/main/DC [accessed 2023-11-01]

Abbreviations

A1C: glycated hemoglobin
ALK FISH: anaplastic lymphoma kinase fluorescence in situ hybridization
ALK IHC: anaplastic lymphoma kinase immunohistochemistry
AUC: area under curve
CDF: cumulative distribution function
CS: conditional sampling
CTGAN: conditional tabular generative adversarial network
CopulaGAN: copula generative adversarial network
DC: divide and conquer
DLCO: diffusing capacity of the lungs for carbon monoxide
DT: decision tree
ECOG: Eastern Cooperative Oncology Group
EGFR: epidermal growth factor receptor
FEV1: forced expiratory volume in 1 second
FPR: false-positive rate
FVC: forced vital capacity
GAN: generative adversarial network
KALC-R: Korea Association for Lung Cancer Registry
KSS: Kolmogorov-Smirnov statistic

LGBM: light gradient-boosting machine
ML: machine learning
NSCLC: non–small cell lung cancer
OD: original data
OOB: out-of-bag
RF: random forest
ROC: receiver operating characteristic
RT: radiotherapy
SDG: synthetic data generation
STD: synthetic tabular data
TPR: true-positive rate
TVD: total variation distance
XGBoost: Extreme Gradient Boosting

Edited by C Lovis; submitted 03.04.23; peer-reviewed by DDJ Hwang, C Sun, JL Raisaro; comments to author 23.06.23; revised version received 02.08.23; accepted 28.10.23; published 24.11.23.

Please cite as:
Kang HYJ, Batbaatar E, Choi DW, Choi KS, Ko M, Ryu KS
Synthetic Tabular Data Based on Generative Adversarial Networks in Health Care: Generation and Validation Using the Divide-and-Conquer Strategy
JMIR Med Inform 2023;11:e47859
URL: https://medinform.jmir.org/2023/1/e47859
doi:10.2196/47859
PMID:37999942

Original Paper

A Linked Open Data–Based Terminology to Describe Libre/Free and Open-source Software: Incremental Development Study

Franziska Jahn[1], Dr rer med; Elske Ammenwerth[2], MET, Dr sc hum; Verena Dornauer[2], MSc, Dr med; Konrad Höffner[1], Dr rer nat; Michelle Bindel[2], MSc; Thomas Karopka[3], Dipl-Ing; Alfred Winter[1], Dr sc hum

[1]Institute of Medical Informatics, Statistics and Epidemiology, Faculty of Medicine, Leipzig University, Leipzig, Germany

[2]Institute of Medical Informatics, University for Health Sciences, Medical Informatics and Technology, Hall in Tirol, Austria

[3]GNU Solidario, Las Palmas de Gran Canaria, Spain

Corresponding Author:
Franziska Jahn, Dr rer med
Institute of Medical Informatics, Statistics and Epidemiology
Faculty of Medicine
Leipzig University
Härtelstr. 16-18
Leipzig, 04107
Germany
Phone: 49 341 97 16194
Email: franziska.jahn@imise.uni-leipzig.de

Abstract

Background: There is a variety of libre/free and open-source software (LIFOSS) products for medicine and health care. To support health care and IT professionals select an appropriate software product for given tasks, several comparison studies and web platforms, such as Medfloss.org, are available. However, due to the lack of a uniform terminology for health informatics, ambiguous or imprecise terms are used to describe the functionalities of LIFOSS. This makes comparisons of LIFOSS difficult and may lead to inappropriate software selection decisions. Using Linked Open Data (LOD) promises to address these challenges.

Objective: We describe LIFOSS systematically with the help of the underlying Health Information Technology Ontology (HITO). We publish HITO and HITO-based software product descriptions using LOD to obtain the following benefits: (1) linking and reusing existing terminologies and (2) using Semantic Web tools for viewing and querying the LIFOSS data on the World Wide Web.

Methods: HITO was incrementally developed and implemented. First, classes for the description of software products in health IT evaluation studies were identified. Second, requirements for describing LIFOSS were elicited by interviewing domain experts. Third, to describe domain-specific functionalities of software products, existing catalogues of features and enterprise functions were analyzed and integrated into the HITO knowledge base. As a proof of concept, HITO was used to describe 25 LIFOSS products.

Results: HITO provides a defined set of classes and their relationships to describe LIFOSS in medicine and health care. With the help of linked or integrated catalogues for languages, programming languages, licenses, features, and enterprise functions, the functionalities of LIFOSS can be precisely described and compared. We publish HITO and the LIFOSS descriptions as LOD; they can be queried and viewed using different Semantic Web tools, such as Resource Description Framework (RDF) browsers, SPARQL Protocol and RDF Query Language (SPARQL) queries, and faceted searches. The advantages of providing HITO as LOD are demonstrated by practical examples.

Conclusions: HITO is a building block to achieving unambiguous communication among health IT professionals and researchers. Providing LIFOSS product information as LOD enables barrier-free and easy access to data that are often hidden in user manuals of software products or are not available at all. Efforts to establish a unique terminology of medical and health informatics should be further supported and continued.

(JMIR Med Inform 2023;11:e38861) doi:10.2196/38861

KEYWORDS

health informatics; ontology; free/libre open-source software; software applications; health IT; terminology

Introduction

Background

Libre/free and open-source software (LIFOSS) products are increasingly used to support various tasks in health care. LIFOSS generally refers to software products with openly available source code that users and developers can view, analyze, modify, and redistribute.

For example, there are free and open-source software products to implement radiology information systems, picture archiving and communication systems (PACS), patient administration systems, and electronic health record (EHR) systems. Especially in low-resource settings, using LIFOSS can help establish computer-based health information systems [1-3]. Along with their use in hospital and medical practice settings, LIFOSS products are available for mobile health, telemedicine, and public health (eg, [4,5]). Moreover, the COVID-19 pandemic has led to the development of numerous mobile applications for contact tracing, risk assessment, or appointment scheduling, which are often based on LIFOSS and used both in low-resource settings and industrial countries [6,7]. Since 2010, the Medfloss.org database has provided descriptions of LIFOSS used in health care and medicine [8]. As of October 2022, it lists 385 software products and describes them by characteristics like "license," "application type," "enterprise function," "language," "platform," and "home page." Due to the iterative and sometimes uncontrolled growth of the self-developed nomenclature over the last few years, there are several inconsistencies in the software descriptions on Medfloss.org. First, there are misleading assignments of descriptors to characteristics. For example, "laboratory," "cellular networks," and "virtual reality" are listed as enterprise functions supported by a software product. However, they describe the setting where the software might be used or special features of the software. Second, the lack of uniform terminology in health informatics has led to the use of synonyms or overlapping terms. For example, the borders between "electronic health systems," "electronic medical record systems," and "hospital management systems" are not clearly defined, sometimes leading to ambiguous descriptions of software products. The lack of uniform terminology for describing medical and health care software products and LIFOSS is also apparent when analyzing comparisons of LIFOSS for EHR systems. In several studies published during the last 15 years [1-3,9-11], each research group selected different criteria and descriptors for comparing the technical and functional characteristics of EHR systems.

Indeed, the lack of uniform terminology is not restricted to LIFOSS. There are already several terminologies in health care and health informatics that could potentially be used to describe software. For example, in addition to allowing for precise descriptions of disease patterns, SNOMED (Systematized Nomenclature of Medicine) lists health occupations and environments that may describe the health facilities or settings in which LIFOSS is used. The World Health Organization (WHO) classifies digital health interventions by hierarchies or lists of clients, health care providers, health system managers, data service features, and application system categories [12]. The Health Level 7 (HL7) EHR System Functional Model is a comprehensive catalogue of EHR systems' features [13]. Finally, some health informatics textbooks provide systematic collections of application systems in health care and their typical features (eg, [14]). However, at first sight, these different terminologies are not easily comparable with each other due to the use of synonyms or homonyms. Linking these different terminologies requires semantic analyses based on a uniform set of concepts, along with easy-to-use methods and tools to link these data.

Linked Open Data (LOD) are regarded as the state-of-the-art principle for linking and structuring concepts from different terminologies. LOD are identifiable by a URI and provided using the Resource Description Framework (RDF) standard [15].

Benefits of Unified LIFOSS Terminology Using LOD

A unified LIFOSS terminology using LOD has several advantages. First, the use of predefined and open terminologies supports the search for and comparison of software products. Second, further knowledge, such as results of assessment and evaluation studies, can be linked easily to the software descriptions and thus support evidence-based health informatics. For example, linking systematic descriptions of software products with descriptors from other projects can support ontology-based approaches to software requirements engineering [16].

The Austrian-German research project "Health Information Technology Ontology (HITO)" aims to systematically describe software products and their installations in health care. It uses an underlying ontology named HITO, LOD methods and tools, and freely available catalogues to describe software characteristics. HITO is developed based on different use cases in which precise software descriptions are needed, such as when selecting LIFOSS or commercial software products, searching for evidence about the installation of software products, and communicating about software products among stakeholders in health care. In this paper, we focus on LIFOSS products and describe them with the help of openly available catalogues. Especially for LIFOSS products, information about software characteristics is freely available, and LIFOSS developers are likely to recognize the potential advantage of spreading knowledge about their products with the help of LOD.

Objectives

This study aims to (1) describe LIFOSS systematically with the help of precise descriptors that are captured in HITO and (2) publish HITO and HITO-based software product descriptions as LOD using Semantic Web tools for viewing and querying LIFOSS data on the World Wide Web.

Methods

Requirements Elicitation

Initial Steps

As the first step toward precise descriptions of software products, the classes and relationships that are useful to describe software products had to be identified. HITO was developed and refined iteratively by collecting requirements of several use

cases. Each use case focused on a situation in which a clear terminology for software products was considered essential. These use cases dealt with the description of LIFOSS or commercial software products for potential and current users in health care settings or the description of software product installations in evaluation studies on health IT interventions. We selected diverse use cases to identify the most relevant characteristics for these diverse situations. Based on these use cases, we incrementally built the ontology that contains a general pattern for describing software products in health care.

Use Case 1: Evaluation of Digital Health Interventions

The first HITO use case dealt with evaluation studies in health informatics. In evaluation studies, it is crucial to carefully describe the intervention in a reproducible and clear manner to allow generalizability of the findings and their aggregation, for example, in the form of systematic reviews. We used an inductive approach and extracted software descriptions from 24 randomly selected published health IT evaluation studies [17]. We found that software product installations were mainly described by their features, application system type, organizational units where they are used, and user groups. These characteristics were specified as classes and added to HITO (Table 1). Altogether, the software descriptions found in evaluation studies are sparse and concentrate on selected evaluated features. Evaluation studies also often use inconsistent terms, which motivated us to develop a clear terminology for software products.

Table 1. HITO[a] classes to describe software products in health IT evaluation studies (HITO use case 1).

HITO class (characteristic of software products)	Description and examples
Software product	Piece of software that is sold as a commercial product or distributed under an open-source license
Feature	Functionalities offered by a software product that directly contribute to the fulfillment of 1 or more enterprise functions (eg, email notification of new results, user directory to control any access)
Application system type	Commonly used names for categories of software product installations in health care (eg, radiology information system, CPOE[b] system)
Organizational unit	Health care setting in which the software product is used and an evaluation study was conducted (eg, laboratory, department of pediatrics)
User group	Health care staff who uses the software product installation (eg, nurse, radiologist)

[a]HITO: Health Information Technology Ontology.

[b]CPOE: computerized physician order entry.

Use Case 2: Description of LIFOSS With Medfloss.org Project Database

In the LIFOSS use case, we extended the list of HITO classes by carefully analyzing the Medfloss.org project database. Medfloss.org aims to offer an overview of LIFOSS projects related to medical informatics and health care [8]. Although Medfloss.org is not maintained anymore, it is still provided in cooperation with 3 LIFOSS-related working groups of the International Medical Informatics Association, the European Federation for Medical Informatics Association, and the International Society for Telemedicine and eHealth. Within Medfloss.org, the LIFOSS products are described by using a predefined set of categories.

We started this use case by surveying 2 operators of the Medfloss.org database. They were asked independently to answer a survey with 11 open questions and 1 closed question. The survey asked about the users of Medfloss.org, the relevance of the categories used to describe software products, and the positive and negative experiences with the categories used to describe LIFOSS.

The results of this survey show that Medfloss.org is intended to be used by physicians or other health care staff, IT administrators, information managers, and software developers to select appropriate LIFOSS for certain health care tasks. Each LIFOSS product is described by 11 categories, such as “enterprise function,” “application type,” and “license,” on the platform. For each category, the list of descriptors has grown over the years as new LIFOSS descriptions were added.

The answers of the platform operators dealing with the assessment of the current terminology and its representation on the platform were arranged according to a SWOT (Strengths, Weaknesses, Opportunities, and Threats) analysis (Textbox 1).

Both operators rated the “enterprise function,” “application type,” “status,” “license,” “standard,” “language,” “client type,” and “platform” categories of Medfloss.org terminology as “important” or “very important” for the description of LIFOSS. The categories “popularity,” “database,” and “programming language/toolkit” were rated less important by 1 of the website operators.

The strengths of the Medfloss.org terminology used include the rough categorization of software products by application system types and the faceted search on the website that is based on categories for describing the software. However, the survey showed that the categories currently lack the possibility to describe the functionalities and application types with enough precision. Therefore, an opportunity to support Medfloss.org terminology may be to integrate existing terminologies such as WHO’s classification of digital health interventions [12] or the HL7 EHR System Functional Model [13]. Furthermore, the folksonomy (ie, the collection of users’ tags for certain objects) of the platform users is not handled by the search functions on Medfloss.org [18].

Based on the SWOT analysis, we selected 9 (out of 10) of the Medfloss.org categories and added these to HITO (Table 2). Some classes used on Medfloss.org were renamed, such as "standard," which was changed to "interoperability standard" to sum up interoperability standards, as well as frameworks describing how to use interoperability standards, such as Integrating the Healthcare Enterprise (IHE). A more fine-grained classification according to interoperability levels was examined for common interoperability standards but proved to be impractical. Many interoperability standards such as HL7 Fast Healthcare Interoperability Services (FHIR) can be assigned to multiple interoperability levels. The Medfloss.org category "programming language/toolkit" was split into 2 classes to distinguish between these different concepts.

Textbox 1. SWOT (Strengths, Weaknesses, Opportunities, and Threats) analysis for Medfloss.org.

Strengths

- Current set of categories to describe libre/free and open-source software
- Rough categorization by application system types
- Usefulness of the categories to provide a faceted search on the platform

Weaknesses

- Conceptual overlaps in categories (eg, electronic health record and electronic medical record)
- Missing hierarchies for enterprise functions
- Missing detailed functional descriptions

Opportunities

- Enhancements of categories by existing terminologies seems possible
- Modeling of user group–dependent categories may increase usefulness

Threats

- No handling of synonyms within users' search terms

Table 2. HITO[a] to describe LIFOSS[b] products (HITO use case 2).

Medfloss.org class name	HITO class	Description and examples
Client type	Client	The client type on which a software product can be run (mobile, native, or web).
Database	Database management system	Some examples are PostgreSQL[c] or MySQL.
Enterprise function	Enterprise function	Describes what action humans or machines must carry out in a certain enterprise to contribute to its mission or goals (eg, patient admission, order entry).
Home page	Home page	Home page of the software product or its development project.
Standard	Interoperability standard	Ability of 2 or more components to exchange information and to use the information that has been exchanged. Under this class name, interoperability standards (eg, HL7[d] FHIR[e] or DICOM[f]) or frameworks describing how to use standards (eg, IHE[g]) are summed up.
Language	Language	Languages in which the software product is available (eg, English, French, and German).
License	License	The license under which a software product is distributed.
Platform	Operating system	The operating system a software uses (eg, Windows). A software product might be able to run on a variety of operating systems.
Programming language/toolkit	Programming language	The programming language used to develop a software product (eg, Java or Python).
Programming language/toolkit	Programming library or toolkit	Programming toolkits are utility programs that are used to develop and maintain software. Programming libraries are a collection of prewritten functions that are ready to be used in coding. Both help programmers develop software in a fast and safe manner.

[a]HITO: Health Information Technology Ontology.

[b]LIFOSS: libre/free and open-source software.

[c]SQL: Structured Query Language.

[d]HL7: Health Level 7.

[e]FHIR: Fast Healthcare Interoperability Resources.

[f]DICOM: Digital Imaging and Communications in Medicine.

[g]IHE: Integrating the Healthcare Enterprise.

Further HITO Use Cases

We summarized 3 further use cases that have not elicited new HITO classes related to software product characteristics.

The third use case deals with the description of commercial software products used in health care. This use case confirmed the set of HITO classes that we already identified by describing LIFOSS products. However, for commercial software products, it is quite challenging to describe them based on these classes because meaningful descriptions of commercial software products are rarely publicly available.

In the fourth use case, the existing HITO classes were linked with competency levels for IT staff in health care organizations, which might be useful for creating job advertisements.

In the fifth use case, findings from the HITO project were discussed with practitioners like hospital chief information officers and industry representatives to discuss the applicability and broader use of the HITO project's findings in practice.

Software Product Descriptions as LOD

After building HITO from the described use cases, we published HITO and HITO-based software product descriptions using LOD.

LOD are web data with an open license. They allow the use of Semantic Web tools for viewing and querying LIFOSS data on the World Wide Web. LOD also allow for linking and reusing existing terminologies. To be considered "5-star LOD," the data need to be machine-readable, presented in a nonproprietary format, use open standards of the World Wide Web Consortium (W3C), and be linked to other data [15]. We strived to achieve 5-star linked open HITO data.

Accordingly, we used RDF Schema (RDFS) and the Web Ontology Language (OWL). RDFS and OWL are W3C standards used to define the types of elements of discourse as classes. The *class* "software product," for example, represents the set of all individual software products. *Properties* represent possible binary-typed relationships between individuals of certain classes, for example, between software products and features. Using RDF, facts (relationships between individuals) are expressed as subject-predicate-object triples, whose elements may be defined and stored in different places. This allows for reusing existing vocabularies and interlinking with existing knowledge bases, forming the LOD cloud. Each RDF resource (a class, an individual, or a relationship) has a URL where it is published both in human-readable (ie, HTML) format and in machine-processable RDF serialization format. To browse RDF data comfortably, tools like RickView [19] can be used and modified. SPARQL Protocol and RDF Query Language

(SPARQL) end points allow free access to structured read-only queries for humans and as application programming interfaces (APIs) for several tools.

Integration of Software Product and Health-Related Terminologies

One advantage of LOD is their easy integration of existing data sources that are already available in RDF format. Therefore, for the HITO classes (Tables 1 and 2) that characterize software products, we searched for "catalogues" (ie, lists of terms that can be used as instances for the respective class). For the selection of suitable catalogues, we defined the following criteria:

[The catalogue should be authored by an established scientific or standardization organization.] OR [The catalogue must be scientifically plausible with regard to reproducibility, having undergone peer-review or having been developed by more than 5 persons.] OR [The catalogue must be openly available and developed by a large community.]

We used a broad literature and web search and our knowledge of the field to identify related taxonomies and catalogues that can be useful for describing instances of HITO classes (Table 2). In the following paragraphs, we provide a brief overview of the catalogues we investigated.

We started with DBpedia, a popular and large knowledge base comprising billions of triples extracted from Wikipedia, texts, and other sources [20,21]. We analyzed DBpedia to identify possible instances or subclasses for the HITO classes. For the classes "language," "operating system," and "programming language," we found DBpedia classes with suitable instances that we replaced the associated HITO classes with to increase interoperability. Other instances of DBpedia classes, such as the "license" class, were not suitable for integration because DBpedia does not semantically differentiate between licenses for software products and licenses for other purposes, like drivers' licenses. For software product licenses, we integrated subclasses of the class "open-source software license" derived from the Software Ontology for biomedical software [22].

The most challenging task was the integration of catalogues for the classes "application system type," "enterprise function," "feature," "organizational unit," and "user group," for which we needed instances or subclasses related to health care. We checked the following sources for the integration of catalogues into HITO:

- HL7 EHR System Functional Model [13]: Using the examples of 2 installations of commercial software products for EHR systems, we assessed whether the features of the software product could be described by this model. We found that the whole list of conformance criteria defined in this model would be too detailed for a HITO catalogue. However, the section labels that are provided by this model, such as "manage allergy, intolerance, and adverse reaction list," provide an appropriate level of detail for feature descriptions in HITO.
- The textbook by Winter et al [14] describes a set of application system types and a set of enterprise functions in hospitals. An analysis of Medfloss.org revealed that the sets of enterprise functions or application system types could be used to tag 71% and 42%, respectively, of 356 Medfloss.org software products analyzed in use case 2 [23]. Due to the textbook's focus on hospitals, appropriate terms for software products used in other health institutions or for public health were missing.
- Taxonomy for health IT by Varshney et al [24]: The application system types identified by Varshney and colleagues proved to be too coarse-grained for classification. Only 105 (29%) of 356 Medfloss.org software product entries could be classified by the taxonomy [23].
- WHO classification of digital health interventions [12]: This multiaxial classification lists system categories and interventions for clients, health care providers, health system managers, and data services. The 25 system categories correspond to application system types in HITO, and their strength lies in their focus on application systems for health care networks rather than for single institutions. For the hierarchically grouped interventions such as "2.10 laboratory and diagnostics imaging management" and "2.10.1 transmit diagnostic result to healthcare provider," a more differentiated assignment to functions and features in the context of HITO is needed. Therefore, "2.10 laboratory and diagnostics imaging management" is listed as an enterprise function in HITO, whereas "2.10.1 transmit diagnostic result to healthcare provider" is listed as a feature in HITO.
- SNOMED Clinical Terms (CT) [25]: For the 24 evaluation studies analyzed in use case 1, we found that the software products' user groups or the organizational units where the software products are used can sufficiently be described by subclasses of SNOMED CT's "occupation" or "environment" classes, respectively.
- Features for PACS selection [26]: The PACS features described on 3 hierarchy levels were transformed into a flat list of 38 features useful for describing examples of PACS software products.
- The LIS (laboratory information system) Functionality Assessment Toolkit of the Association for Pathology Informatics [27]: This toolkit lists 850 functionality statements and describes a methodology for selecting a LIS. In comparison to the other feature catalogues we described, the functionality statements are very detailed. For integration into HITO, a clustering of the functionality statements would be necessary to obtain a manageable LIS feature list.

Describing Software Products Using HITO

With the help of HITO and the selected catalogues we described, 25 LIFOSS products were described. These were selected to represent different application system types and due to their comprehensive, openly available documentation. For the description of supported enterprise functions and features, we extracted terms from the software product documentation and linked as many as possible to catalogue entries of HITO. Information about the LIFOSS products was extracted by 1 team member (author MB) and checked by another team member (author FJ).

Results

HITO was modeled and published using Semantic Web technologies. The Unified Modeling Language class diagram in Figure 1 describes the structure of HITO. The classes shown to the left of "software product" describe the general characteristics of software products. For classes with domain-specific instances (ie, application system type, feature, enterprise function, organizational unit, and user group), we applied a scheme of 3 interrelated classes named <*classname*>Catalogue, <*classname*>Classified, and <*classname*>Citation. A <*classname*>Catalogue is a health IT–related collection, such as the HL7 EHR Functional Model or the WHO classifications for digital interventions. A <*classname*>Classified is a category that belongs to exactly 1 catalogue. A <*classname*>Citation is a textual label extracted from available software manuals, descriptions, and studies and thus represents a part of the folksonomy contained in HITO [28].

To illustrate how this scheme applies, we describe the domain-specific characteristics of the Orthanc software [29,30] in Table 3. The developers of Orthanc refer to it as "mini-PACS," "DICOM (Digital Imaging and Communications in Medicine) server," "VNA (vendor-neutral archive)," and "viewer of medical images." These terms are assigned to the software product Orthanc as "application system type citations." In turn, "DICOM server" and "mini-PACS" have links to the classified application system type "PACS" from the application system type catalogue in Winter et al [14]. The supported enterprise function citations extracted from the Orthanc website, such as "image archiving," "image management," and "research about the automated analysis of medical images," were linked to the more general terms "laboratory and diagnostics imaging management" and "research and education" from enterprise function catalogues. For the 16 feature citations extracted from the Orthanc online documentation, we identified linkable classified features of the WHO classification of digital health interventions and the PACS feature list. Some classified features or enterprise function terms have direct links to the software product. These assignments had no match with citations and were done by domain experts.

Overall, with the help of HITO, we described 25 LIFOSS products in similar detail as Orthanc. All software product descriptions are available as LOD. We described HITO classes and relationships of the ontology and individual software products using RDF. The HITO SPARQL end point [31] allows queries using SPARQL (Figure 2).

The RickView application allows for browsing through the ontology and knowledge base [32]. Another suitable way to query is with a faceted search [33], whereby integrated terminologies can be used to find software products of a certain application system type or supporting combinations of features and functions (Figure 3).

The ontology is also publicly available under version control in a GitHub repository [34]. It can be downloaded in RDF Turtle format to be viewed in ontology editors like Protégé. HITO is dedicated to the public domain and uses a Creative Commons Zero v1.0 Universal license. There are a few exceptions for integrated terms from SNOMED [25] and the WHO classification of digital health interventions [12].

Figure 1. The Health Information Technology Ontology (HITO), version 22.05, specifies the classes and relationships that are used to describe software products. The complete class diagram is available on the HITO website [28].

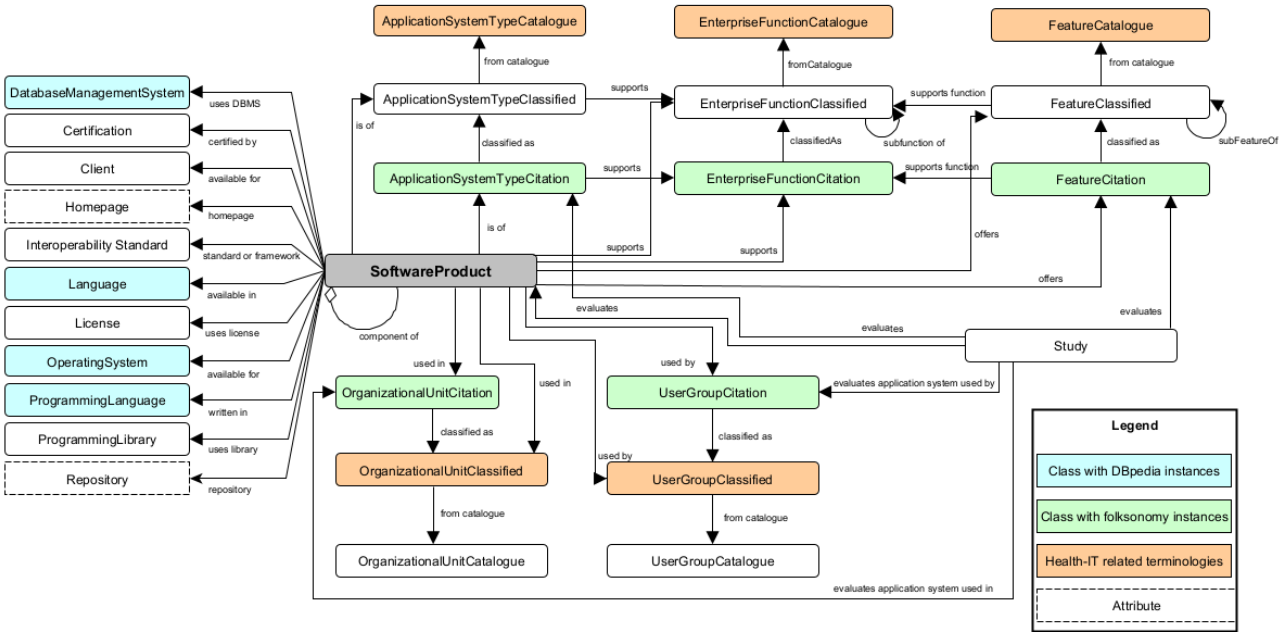


Table 3. Application system types, supported functions, features, user groups and organizational units of the Orthanc software.

<className> and <classname>Citation from the software documentation	<classname>Classified in HITO[a] (<classname>Catalogue)
Application system type	
DICOM[b] server	PACS[c] (application systems in [14])
mini-PACS	PACS (application systems in [14])
VNA[d]	N/A[e]
Viewer of medical images	N/A
Web viewer	N/A
Enterprise function	
N/A	Execution of radiological examinations (enterprise functions from [14])
Image archiving	Laboratory and diagnostics imaging management (enterprise functions from [12])
Image communication	Laboratory and diagnostics imaging management (enterprise functions from [12])
Image distribution	Laboratory and diagnostics imaging management (enterprise functions from [12])
Image management	Laboratory and diagnostics imaging management (enterprise functions from [12])
Research about the automated analysis of medical images	Research and education (enterprise functions from [14])
Feature	
N/A	Capture diagnostic results from digital devices (features from [12])
Data management for clinical routine and medical research	N/A
DICOM identifiers	N/A
DICOM network protocol	N/A
Evaluations Report	N/A
Injury surveillance system registration report	N/A
Listing available servers	N/A
Plugin mechanism to add new modules	N/A
Retrieve images	N/A
Retrieving DICOM resources from WADO-RS[f] server	Compatibility and integration with other systems and products (PACS feature list [26])
Search the content	N/A
Send images	N/A
Sending DICOM resources to a STOW-RS[g] server	Compatibility and integration with other systems and products (PACS feature list [26])
Test the connection	N/A
Top diseases report	N/A
User group	
Radiologist	Radiologist occupation [25]
Researcher	Researcher occupation [25]
Software/hardware integrators in the medical field	N/A
Network engineer	N/A
Physicist	Physicist occupation [25]
System engineer	N/A
Organizational unit	
Health centers	Health center environment [25]
Hospital environment	Hospital environment [25]

<className> and <classname>Citation from the software documentation	<classname>Classified in HITO[a] (<classname>Catalogue)
Radiology department	Radiology department environment [25]

[a]HITO: Health Information Technology Ontology.

[b]DICOM: Digital Imaging and Communications in Medicine.

[c]PACS: picture archiving and communication system.

[d]VNA: vendor-neutral archive.

[e]N/A: not applicable.

[f]WADO-RS: Web Access to DICOM Objects by RESTful (representational state transfer) Services.

[g]STOW-RS: Store Over the Web by RESTful Services.

Figure 2. A SPARQL Protocol and RDF Query Language (SPARQL) query (on the left) and its results (on the right). RDF: Resource Description Framework.

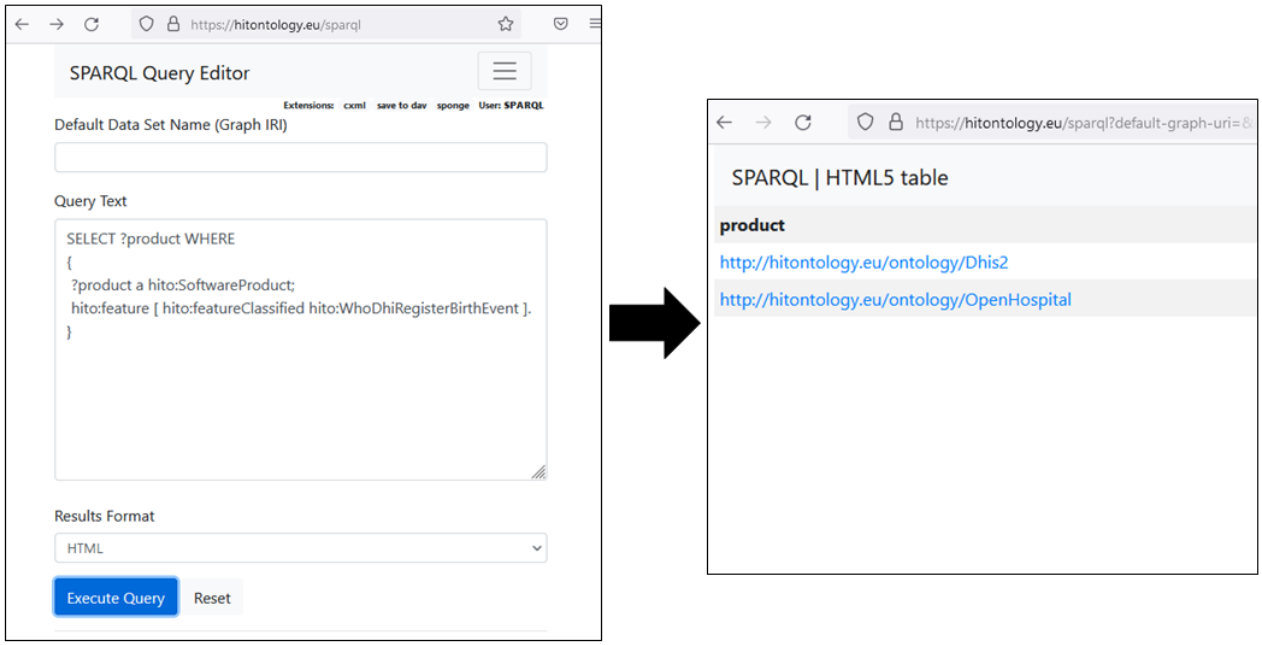
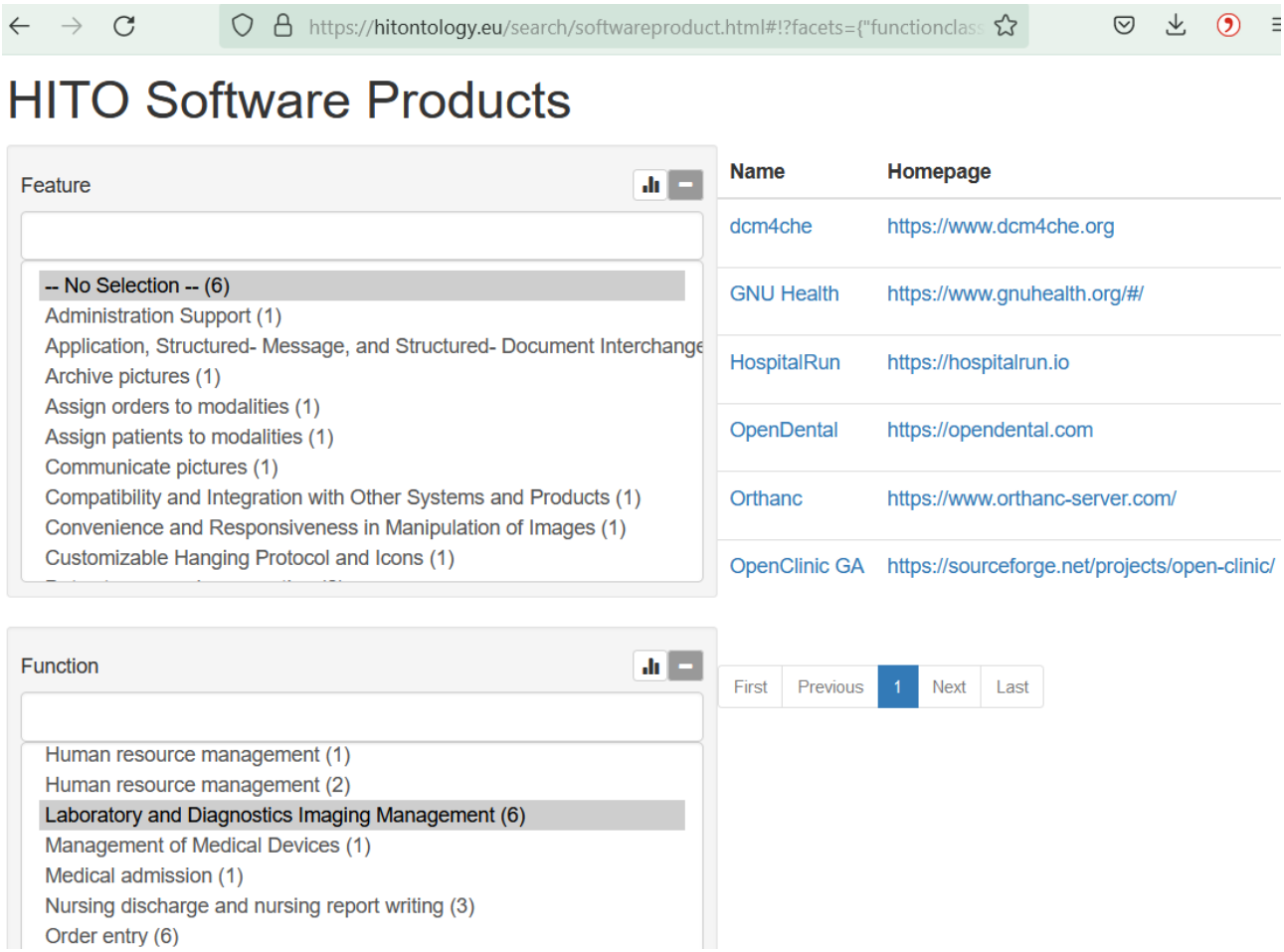


Figure 3. Faceted search for software products supporting "Laboratory and Diagnostics Imaging Management". HITO: Health Information Technology Ontology.



Discussion

Principal Results

In this project, we identified characteristics useful for describing software products and systematically captured them as classes in HITO. Accordingly, we exploited the properties of ontologies that enable semantic description and linking of data.

For a thorough functional description of software products in health care, we described the enterprise functions supported and features offered by the software products. For health-specific characteristics of software products, we analyzed and integrated existing terminologies for enterprise functions, features, application system types, organizational units, and user groups as catalogues. Because we expected the relevant sources for such catalogues to come not only from science but also from practice, a systematic review of the scientific literature for health IT–related terminologies would not have led to sufficient results. Accordingly, we based the selection of catalogues on statements by domain experts among the project team and project partners, supplemented by targeted PubMed and Google searches on specific application system types.

Thus, we used a case-based and agile approach to identify classes, relationships, and catalogues best suited for describing software products in health care. In the use cases considered so far, ontological reasoning had no relevance. Therefore, to date, only few axioms are used in HITO. As the next step toward more interoperability with other formal ontologies, HITO could use an upper-level ontology such as Basic Formal Ontology [35], General Formal Ontology [36], or gist [37]. A first feasibility check of these ontologies showed that the gist ontology, which defines typical upper-level business concepts, may be the most appropriate for the scope of HITO.

With the help of HITO, we described 25 LIFOSS products in detail that, together with less detailed descriptions of single commercial software products and software products extracted from evaluation studies, form our knowledge base. The descriptions of these software products could be regarded as a proof of concept. However, we noticed the interpretative degrees of freedom in assigning correct enterprise functions and features to software products. To ensure the validity of further software descriptions, it would be helpful to calculate interrater reliability among 2 independent experts. For this, further software product entries of the Medfloss.org database that have not yet been considered in HITO's knowledge base could be used.

As postulated by Berners-Lee [15], HITO's availability as LOD facilitates its barrier-free access and use. In particular, the integrated catalogues for enterprise functions, features, and application system types provide HITO users with rich terminology for functionalities of software products. However, since there is more than 1 catalogue for each of these characteristics, new terminological problems arise. The

catalogue entries of different catalogues must be mapped to each other to achieve comparability of software products described with the help of different catalogues. Linking these catalogues is part of ongoing research. Together with the folksonomy terms that are already connected to catalogue entries, HITO users will be able to retrieve the most suitable software using a broad range of search terms. The catalogues currently integrated into HITO focus more on health care rather than on medical research tasks (ie, software products like research databases may not be sufficiently described by HITO). However, integrating further catalogues describing research-related enterprise functions or features could be possible.

Nevertheless, publishing HITO and its knowledge base as LOD implies that the contents of HITO are available under an open license. Thus, for the broadly accepted nomenclature SNOMED CT, we could only check the principal suitability of the SNOMED "environment" and "occupation" concepts based on a small set of examples that we included in HITO with permission from SNOMED International. Due to license requirements, SNOMED CT terms cannot be made available as LOD.

In summary, HITO provides an openly available framework for the description of health care–related software that can be used by researchers who publish studies on digital health interventions or by developers and users who need to describe software.

Conclusions

We recognize that health informatics continues to face a terminology problem. Establishing a uniform terminology for software products used in health care is currently unachievable due to several coexisting terminologies from both research and practice. Linking the terms from different terminologies by similarity relationships is the first step toward more transparency. This will also help identify misunderstandings that may be caused by synonyms, homonyms, or conceptual overlaps. Simply knowing that the term "EHR system" can stand for an institutional or cross-institutional application system or a collection of digital documents related to a person's health prevents problems related to misunderstanding. A researcher authoring a study on a digital health intervention by an EHR system knows that the term "EHR system" must be further specified, for example, by enterprise functions and features as listed in HITO.

Nevertheless, we should further strive for a consented, uniform terminology of health informatics. Taking different coexisting terminologies as a basis, methods of qualitative content analyses such as inductive category formation [38], supported by (semi)automatic text extraction, may lead the way toward an established language for health informatics.

Acknowledgments

The work presented here is part of the project "HITO: a Health IT Ontology," funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - WI 1605/11-1 and the Austrian Science Fund (I 3726-N31). The publication of this article was partly funded by the Open Access Publishing Fund of Leipzig University, supported by the German Research Foundation for the Open Access Publication Funding program.

Conflicts of Interest

None declared.

References

1. Aminpour F, Sadoughi F, Ahamdi M. Utilization of open source electronic health record around the world: A systematic review. J Res Med Sci 2014 Jan;19(1):57-64 [FREE Full text] [Medline: 24672566]
2. Syzdykova A, Malta A, Zolfo M, Diro E, Oliveira JL. Open-source electronic health record systems for low-resource settings: systematic review. JMIR Med Inform 2017 Nov 13;5(4):e44 [FREE Full text] [doi: 10.2196/medinform.8131] [Medline: 29133283]
3. Millard PS, Bru J, Berger CA. Open-source point-of-care electronic medical records for use in resource-limited settings: systematic review and questionnaire surveys. BMJ Open 2012;2(4) [FREE Full text] [doi: 10.1136/bmjopen-2011-000690] [Medline: 22763661]
4. Chen C, Haddad D, Selsky J, Hoffman JE, Kravitz RL, Estrin DE, et al. Making sense of mobile health data: an open architecture to improve individual- and population-level health. J Med Internet Res 2012;14(4):e112 [FREE Full text] [doi: 10.2196/jmir.2152] [Medline: 22875563]
5. Ngabo F, Nguimfack J, Nwaigwe F, Mugeni C, Muhoza D, Wilson DR, et al. Designing and implementing an innovative SMS-based alert system (RapidSMS-MCH) to monitor pregnancy and reduce maternal and child deaths in Rwanda. Pan Afr Med J 2012;13:31 [FREE Full text] [Medline: 23330022]
6. Rawat S. Impact of open source during COVID-19 pandemic. OpenSense Labs. 2021 Jan 28. URL: https://opensenselabs.com/blog/articles/open-source-covid-19-pandemic [accessed 2022-12-20]
7. Alanzi T. A review of mobile applications available in the app and Google Play stores used during the COVID-19 outbreak. J Multidiscip Healthc 2021;14:45-57 [FREE Full text] [doi: 10.2147/JMDH.S285014] [Medline: 33469298]

8. Streidl H, Demski H, Karopka T. Medical free/libre and open source software. Medfloss.org. URL: https://medfloss.org/ [accessed 2022-01-13]
9. Zaidan AA, Zaidan BB, Al-Haiqi A, Kiah MLM, Hussain M, Abdulnabi M. Evaluation and selection of open-source EMR software packages based on integrated AHP and TOPSIS. J Biomed Inform 2015 Mar;53:390-404 [FREE Full text] [doi: 10.1016/j.jbi.2014.11.012] [Medline: 25483886]
10. Purkayastha S, Allam R, Maity P, Gichoya JW. Comparison of open-source electronic health record systems based on functional and user performance criteria. Healthc Inform Res 2019 Apr;25(2):89-98 [FREE Full text] [doi: 10.4258/hir.2019.25.2.89] [Medline: 31131143]
11. Flores ZAE, Win KT, Susilo W. Functionalities of free and open electronic health record systems. Int J Technol Assess Health Care 2010 Oct;26(4):382-389. [doi: 10.1017/S0266462310001121] [Medline: 20974022]
12. WHO. Classification of Digital Health Interventions v 1.0: a shared language to describe the uses of digital technology for health. World Health Organization. 2018. URL: https://apps.who.int/iris/bitstream/handle/10665/260480/WHO-RHR-18.06-eng.pdf [accessed 2022-12-22]
13. HL7 Electronic Health Record System Functional Model, Release 2.1. Health Level 7 International. 2020. URL: http://www.hl7.org/implement/standards/product_brief.cfm?product_id=528 [accessed 2022-12-21]
14. Winter A, Haux R, Ammenwerth E, Brigl B, Hellrung N, Jahn F. Health Information Systems - Architectures and Strategies. London, UK: Springer; 2011.
15. Berners-Lee T. Linked data: design issues. W3C. 2009. URL: https://www.w3.org/DesignIssues/LinkedData.html [accessed 2019-11-06]
16. Dermeval D, Vilela J, Bittencourt II, Castro J, Isotani S, Brito P, et al. Applications of ontologies in requirements engineering: a systematic review of the literature. Requirements Eng 2015 Feb 14;21(4):405-437. [doi: 10.1007/s00766-015-0222-6]
17. Dornauer V, Jahn F, Hoeffner K, Winter A, Ammenwerth E. Use of natural language processing for precise retrieval of key elements of health IT evaluation studies. Stud Health Technol Inform 2020 Jun 26;272:95-98. [doi: 10.3233/SHTI200502] [Medline: 32604609]
18. Trant J. Studying social tagging and folksonomy: a review and framework. Digital Libraries and User-Generated Content 2009;10(1):1-44.
19. Höffner K. RickView: A fast RDF viewer (linked data browser). Crates.io. 2022. URL: https://crates.io/crates/rickview [accessed 2022-10-18]
20. Lehmann J, Isele R, Jakob M, Jentzsch A, Kontokostas D, Mendes P, et al. DBpedia - A large-scale, multilingual knowledge base extracted from Wikipedia. SWJ 2015;6(2):167-195. [doi: 10.3233/sw-140134]
21. Hofer M, Hellmann S, Dojchinovski M, Frey J. The new DBpedia release cycle: increasing agility and efficiency in knowledge extraction workflows. 2020 Presented at: 16th International Conference on Semantic Systems; September 7–10; Amsterdam, the Netherlands p. 1-18. [doi: 10.1007/978-3-030-59833-4]
22. Malone J, Brown A, Lister AL, Ison J, Hull D, Parkinson H, et al. The Software Ontology (SWO): a resource for reproducibility in biomedical data analysis, curation and digital preservation. J Biomed Semantics 2014;5:25 [FREE Full text] [doi: 10.1186/2041-1480-5-25] [Medline: 25068035]
23. Jahn F, Bindel M, Höffner K, Ghalandari M, Schneider B, Stäubert S, et al. Towards precise descriptions of medical free/libre and open source software. Stud Health Technol Inform 2020 Jun 16;270:463-468. [doi: 10.3233/SHTI200203] [Medline: 32570427]
24. Varshney U, Nickerson RC, Muntermann J. Taxonomy development in health-IT. In: Proceedings of the 19th Americas Conference on Information Systems. 2013 Presented at: AMCIS'13; August 15-17; Chicago, Illinois p. 1-10.
25. SNOMED CT browser. SNOMED International. URL: https://browser.ihtsdotools.org/? [accessed 2022-01-19]
26. Joshi V, Narra VR, Joshi K, Lee K, Melson D. PACS administrators' and radiologists' perspective on the importance of features for PACS selection. J Digit Imaging 2014 Aug;27(4):486-495 [FREE Full text] [doi: 10.1007/s10278-014-9682-3] [Medline: 24744278]
27. LIS Functionality Assessment Toolkit. Association for Pathology Informatics. URL: https://www.pathologyinformatics.org/lis-toolkit [accessed 2022-01-19]
28. HITO: A Health IT ontology for systematically describing application systems and software products in health it. HITO. 2022. URL: https://hitontology.eu [accessed 2022-04-12]
29. Jordogne S, Osimis SE. Orthanc: Open-source, lightweight DICOM server. Orthanc. 2022. URL: https://www.orthanc-server.com/index.php [accessed 2022-02-07]
30. Jodogne S. The Orthanc ecosystem for medical imaging. J Digit Imaging 2018 Jun;31(3):341-352 [FREE Full text] [doi: 10.1007/s10278-018-0082-y] [Medline: 29725964]
31. SPARQL Query Editor. HITO. URL: http://hitontology.eu/sparql [accessed 2022-03-24]
32. HITO Health IT Ontology. HITO. URL: https://hitontology.eu/ontology/ [accessed 2022-03-22]
33. HITO software products search. HITO. URL: https://hitontology.eu/search/softwareproduct.html [accessed 2022-03-24]
34. The Health IT Ontology. GitHub Repository. URL: https://github.com/hitontology/ontology [accessed 2022-10-13]
35. 21838-2:2021 Information technology — Top-level ontologies (TLO) — Part 2: Basic Formal Ontology (BFO). International Organization for Standardization. 2021. URL: https://www.iso.org/standard/74572.html [accessed 2022-10-26]

36. Loebe F, Burek P, Herre H. GFO: the General Formal Ontology. Appl Ontol 2022 Mar 15;17(1):71-106. [doi: 10.3233/ao-220264]
37. Semantic arts: gist. GitHub Repository. URL: https://github.com/semanticarts/gist [accessed 2022-10-25]
38. Mayring P. Qualitative content analysis: theoretical background and procedures. In: Bikner-Ahsbahs A, Knipping C, Presmeg N, editors. Approaches to Qualitative Research in Mathematics Education. Dordrecht, the Netherlands: Springer; 2015:365-380.

Abbreviations

API: application programming interface
DICOM: Digital Imaging and Communications in Medicine
EHR: electronic health record
HITO: Health Information Technology Ontology
HL7: Health Level 7
LIFOSS: libre/free and open-source software
LIS: laboratory information system
LOD: Linked Open Data
OWL: Web Ontology Language
PACS: picture archiving and communication system
RDF: Resource Description Framework
RDFS: Resource Description Framework Schema
SNOMED: Systematized Nomenclature of Medicine
SNOMED CT: Systematized Nomenclature of Medicine Clinical Terms
SPARQL: SPARQL Protocol and RDF Query Language
SWOT: Strengths, Weaknesses, Opportunities, and Threats
VNA: vendor-neutral archive
W3C: World Wide Web Consortium
WHO: World Health Organization

Edited by C Lovis; submitted 19.04.22; peer-reviewed by S Schulz, F Lehocki; comments to author 07.09.22; revised version received 28.10.22; accepted 09.11.22; published 20.01.23.

Please cite as:
Jahn F, Ammenwerth E, Dornauer V, Höffner K, Bindel M, Karopka T, Winter A
A Linked Open Data–Based Terminology to Describe Libre/Free and Open-source Software: Incremental Development Study
JMIR Med Inform 2023;11:e38861
URL: https://medinform.jmir.org/2023/1/e38861
doi:10.2196/38861
PMID:36662569

Review

Systematized Nomenclature of Medicine–Clinical Terminology (SNOMED CT) Clinical Use Cases in the Context of Electronic Health Record Systems: Systematic Literature Review

Riikka Vuokko[1], PhD; Anne Vakkuri[2*], MD, PhD; Sari Palojoki[3*], PhD

[1]Unit for Digitalization and Management, Ministry of Social Affairs and Health, Helsinki, Finland

[2]Perioperative, Intensive Care and Pain Medicine, Helsinki University Hospital, Vantaa, Finland

[3]Unit for Digital Transformation, European Centre for Disease Prevention and Control, Stockholm, Sweden

[*]these authors contributed equally

Corresponding Author:
Riikka Vuokko, PhD
Unit for Digitalization and Management
Ministry of Social Affairs and Health
PO Box 33
Helsinki, FI-00023 Government
Finland
Phone: 358 50 453 6377
Email: riikka.vuokko@gov.fi

Abstract

Background: The Systematized Medical Nomenclature for Medicine–Clinical Terminology (SNOMED CT) is a clinical terminology system that provides a standardized and scientifically validated way of representing clinical information captured by clinicians. It can be integrated into electronic health records (EHRs) to increase the possibilities for effective data use and ensure a better quality of documentation that supports continuity of care, thus enabling better quality in the care process. Even though SNOMED CT consists of extensively studied clinical terminology, previous research has repeatedly documented a lack of scientific evidence for SNOMED CT in the form of reported clinical use cases in electronic health record systems.

Objective: The aim of this study was to explore evidence in previous literature reviews of clinical use cases of SNOMED CT integrated into EHR systems or other clinical applications during the last 5 years of continued development. The study sought to identify the main clinical use purposes, use phases, and key clinical benefits documented in SNOMED CT use cases.

Methods: The Cochrane review protocol was applied for the study design. The application of the protocol was modified step-by-step to fit the research problem by first defining the search strategy, identifying the articles for the review by isolating the exclusion and inclusion criteria for assessing the search results, and lastly, evaluating and summarizing the review results.

Results: In total, 17 research articles illustrating SNOMED CT clinical use cases were reviewed. The use purpose of SNOMED CT was documented in all the articles, with the terminology as a standard in EHR being the most common (8/17). The clinical use phase was documented in all the articles. The most common category of use phases was SNOMED CT in development (6/17). Core benefits achieved by applying SNOMED CT in a clinical context were identified by the researchers. These were related to terminology use outcomes, that is, to data quality in general or to enabling a consistent way of indexing, storing, retrieving, and aggregating clinical data (8/17). Additional benefits were linked to the productivity of coding or to advances in the quality and continuity of care.

Conclusions: While the SNOMED CT use categories were well supported by previous research, this review demonstrates that further systematic research on clinical use cases is needed to promote the scalability of the review results. To achieve the best out-of-use case reports, more emphasis is suggested on describing the contextual factors, such as the electronic health care system and the use of previous frameworks to enable comparability of results. A lesson to be drawn from our study is that SNOMED CT is essential for structuring clinical data; however, research is needed to gather more evidence of how SNOMED CT benefits clinical care and patient safety.

(JMIR Med Inform 2023;11:e43750) doi:10.2196/43750

KEYWORDS

clinical; electronic health record; EHR; review method; literature review; SNOMED CT; Systematized Nomenclature for Medicine; use case; terminology; terminologies; SNOMED

Introduction

Background

The Systematized Medical Nomenclature for Medicine–Clinical Terminology (SNOMED CT) is an extensive, multi-hierarchical clinical terminology system. It provides a standardized and scientifically validated way of representing clinical information [1]. The application possibilities of SNOMED CT are well documented [1-5], and various guides describe the following types of implementation: clinical records, knowledge representation, data aggregation, and analysis. Specifically, the previous literature describes the various development goals of SNOMED CT. For example, SNOMED CT can be used as a standard for electronic health records (EHRs) for classifying or coding clinical information. Additionally, standardized terminology advances data indexing, storing, and retrieving. This supports sharing of patient information across medical domains and organizations in ways that promote continuity of care. As a large-scale terminology system, SNOMED CT also enables knowledge representations in clinical guidelines and care pathways, which can be used, for example, with decision support [2-5].

Data recorded in EHRs are primarily used to provide care to patients. The potential of SNOMED CT to improve data quality and facilitate interoperability, and thus improve patient safety, has long been noted in existing research. Studies have shown that structured and standardized EHRs also increase data reuse possibilities [6]. The European Union and the US Healthcare Information Technology Standards Panel have noted possibilities provided by SNOMED CT and taken steps toward increasing semantic interoperability, reuse, and the exchange of health data. Data recorded in local systems can also be used to support the achievement of broad health policy goals. The importance of SNOMED CT is expected to gradually grow, but at the same time, there is a need to tackle the complex implementation challenges that may arise [1,7,8].

When implemented in EHRs, SNOMED CT is used to represent clinical information consistently and comprehensively [1,2]. Despite the widespread adoption of EHRs that are certified to follow terminology standards, and although SNOMED CT is used in more than 50 countries, there are only a few published reviews about its clinical use. Most studies have focused on theory and predevelopment or design [1,2,8,9]. Moreover, studies in the past have analyzed general factors related to EHR adoption but have not explored the factors associated with less advanced EHR product implementations as compared to more advanced and mature EHR systems. In the context of SNOMED CT implementation, the maturity of an EHR system is a relevant factor to appropriately maximize the benefit from previous experiences [10-12].

In summary, even though SNOMED CT has been extensively studied as a clinical terminology system, previous research has repeatedly documented a lack of detailed evidence for SNOMED CT in clinical use cases [2,3]. Considering that implementing SNOMED CT is a challenging proposition [2], the identification of specific barriers and facilitators to implementing SNOMED CT in clinical use is of paramount importance to further promote its adoption. Evidence from use cases might support implementation and provide guidance on avoiding deployment pitfalls. Therefore, we aimed to explore the available evidence in previous literature reviews of clinical use cases of SNOMED CT integrated into EHR systems or clinical applications during the last 5 years of continued development [2,3,5,13].

Objectives

The aim of this review is to provide an overview of published studies on SNOMED CT clinical use cases in the context of EHRs. In this study, we apply categories from previous research for analyzing use purpose and use phase for the terminology. Moreover, we present core benefits by summarizing the observations from the EHR use cases [3,5,13].

Our research questions are as follows: (1) What are the main clinical use purposes of SNOMED CT during the last 5 years of development? (2) What kinds of use phases of SNOMED CT are identified in these studies? and (3) What are the summarized clinical benefits documented in each SNOMED CT use case?

Methods

To explore EHR-related SNOMED CT use cases in recent research, our research team set out to conduct a systematic literature review. Our team consisted of a medical expert with decades of experience in clinical care and two health and medical informatics experts. To analyze EHR use cases where SNOMED CT terminology was applied, we extended the concept of EHR systems to cover EHR-related applications and software in clinical use. The word “clinical” refers to “medical work or teaching that relates to the examination and treatment of ill people” [14]. In our review, a use case consisted of SNOMED CT integrated into an EHR system in various stages of use, either in preuse development or in design, piloting, testing, implementation, use, or postimplementation evaluation [15].

The team followed the Cochrane review protocol [16] to plan the necessary steps for this study design (Multimedia Appendix 1). Within the team, the application of the protocol was modified step-by-step to fit the research problem by first defining the search strategy, identifying the articles for the review by isolating the exclusion and inclusion criteria for assessing the search results, and lastly evaluating and summarizing the review results.

We defined our search strategy by adding variations of search terms and testing the suitability against the search results. A search with key words “EHR,” “EMR,” “electronic health record,” or “electronic health record system” produced a large number of search results. Combining these terms with

"SNOMED CT" produced relevant search results for our review purposes. Furthermore, adding filters (Textbox 1) did not cause a significant change in the search results. A search of PubMed using the systematic-review methods filter was undertaken in March 2022 and resulted in 162 original articles after removing duplicates (Figure 1). Our results cover the last 5 years of research; thus, our review forms a continuation of previous reviews [2-5].

Textbox 1. Search strategy and filters used.

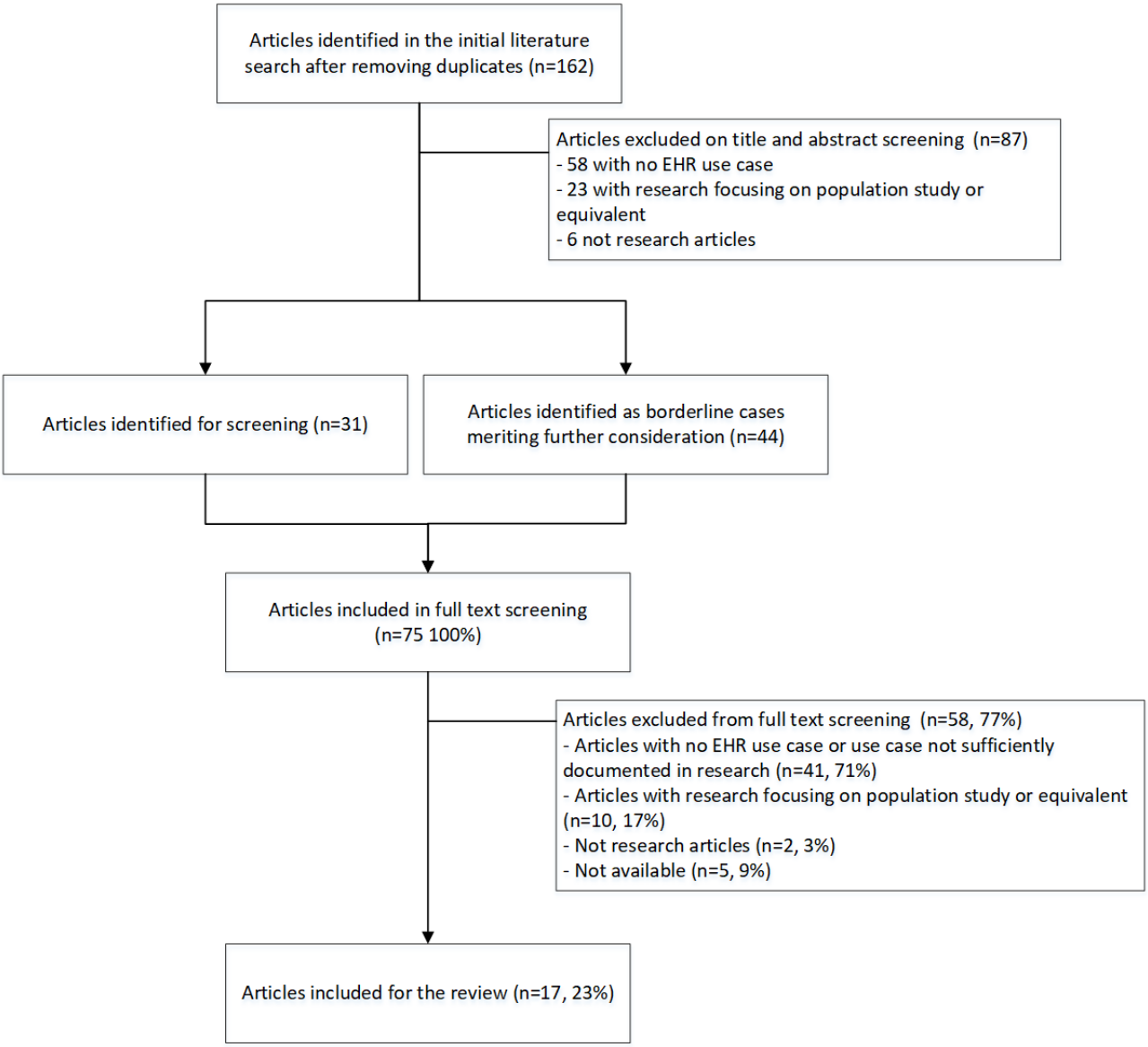
Search terms

(((((Ehr) OR (Emr)) OR (electronic health record)) OR (electronic health record system)) OR (electronic medical record) AND ((fha[Filter]) AND (fft[Filter]) AND (2016/1/1:2022[pdat]))) AND ((Snomed ct) OR (snomed CT) AND ((fha[Filter]) AND (fft[Filter]) AND (2016/1/1:2022[pdat]) AND (english[Filter])))

Filters

Abstract, Full text, English, Abstract, Full text, English, from 2016/1/1 – 2022

Figure 1. Application of the review protocol as a flowchart. EHR: electronic health record.



The exclusion and inclusion criteria were defined before conducting the search. We based our criteria on our research questions and previous research. We defined the exclusion criteria as follows: First, the original article had to document an EHR use case where SNOMED CT was being tested, piloted, implemented, or used in a clinical context. This excluded, for example, research concentrating on theoretical building, evaluation, or validation of SNOMED CT. Second, we excluded population studies and, for example, cohort studies where SNOMED CT was used to define, extract, and harmonize study data and where no EHR-related design or use goals were documented. Third, we excluded editorials, posters, and other

such sources to limit the review to original research articles. While reading the articles, we discussed how well the exclusion criteria corresponded to the delimitation made based on our research questions and the conclusions we arrived at while reading the research content as presented in the original articles.

After the first exclusion based on article headings and abstracts, we had excluded 87 articles. We included 31 articles that seemed to relate to our research question. Moreover, 44 borderline cases merited special consideration to determine if they should be excluded or included (Figure 1). The two researchers set out to read a total of 75 full-text articles. The information extraction and documentation template for final inclusion had been defined based on previous research [3,5,13] and research team agreement (Table 1). During the final reading, the documentation template for information extraction was concurrently refined.

Figure 1 illustrates how the search results were analyzed during exclusion and inclusion screening. In the end, we included 17 original articles in the review analysis. Our final inclusion was confirmed by the research documentation that illustrated the SNOMED CT use case in an EHR or EHR-related application and software in clinical use.

Table 1. Criteria to categorize Systematized Nomenclature of Medicine–Clinical Terminology use in the review.

Criteria	Definition
Clinical use context	Refers to clinical domain or specialty as documented in the study.
EHR[a] system	Refers to EHR systems or other EHR-related applications or software as documented in the study.
Users	Refers to intended users of the SNOMED CT[b] integrated into the EHR as documented in the study.
SNOMED CT use category	Refers to primary purpose for using SNOMED CT as documented in the study [3,5,13]. Based on research team agreement, the following categories were used: standard for EHR or for a clinical application, retrieval or analysis of patient data, data extraction (used to classify or code in a study), proving merit of SNOMED CT, and development of automated coding.
SNOMED CT use phase	Refers to the stage of the SNOMED CT use as documented in the study [3,5]. The stages used by the research team were "in development," "in pilot," "in implementation," "in use," and "after implementation [or in use] evaluation" (ie, proof of merit). Thus, for example, theoretical research was excluded.
SNOMED CT core benefits	Refers to research team's summary of which areas of identified benefits the research added value to, if available [3,5]. The categories were "improving quality of care and patient safety"; "improving continuity of care"; "enabling a consistent way of indexing, storing, retrieving, and aggregating clinical data"; "improving data quality"; and "improving coding productivity."

[a]EHR: electronic health record.

[b]SNOMED CT: Systematized Nomenclature of Medicine–Clinical Terminology.

Results

Characteristics of the Publications

In total, we analyzed 17 articles [17-33]. The earliest research selected for the final review was published in June 2017 and the last in February 2022. On an annual basis, the largest number of articles was published as late as 2021 (6/17). The country of publication was identified for all publications. The country that published the highest number of articles was the United Kingdom (5/17), followed by the United States (3/17), Australia (2/17), and Spain (2/17). Canada, Denmark, Switzerland, the Republic of Korea, and Germany each accounted for one use case. The selected articles were published in 13 different peer-reviewed journals. The characteristics of the publications are summarized in Multimedia Appendix 2.

Contextual Factors of the Clinical Use Cases

To verify the appropriateness of the use cases in regard to our research questions, the clinical context was described for 14 articles. In the remaining 3 articles (3/17), the clinical context was not named, but the research account otherwise described the clinical use case in relation to EHR use. With respect to specialties, 2 cases were from neurology, and 1 case each from pulmonology (asthma), cardiology, oncology, general medicine, pediatrics, and rare diseases. One of the cases did not describe an exact specialty but concerned the prehospital unit in emergency care, and one concerned an outpatient clinic. Four cases mentioned either primary care or tertiary care.

The EHRs were poorly described in most of the articles, and specific descriptions of the EHRs did not follow any uniform structure. Thus, the nature of the results is descriptive. Only one of the publications named the exact product. That study concerned a comprehensive hospital information system, a high-maturity EHR with tools for functions such as supporting care coordination and continuity of care. One of the systems was described as a prehospital patient record that was not integrated into the hospital EHR. One system used by general practitioners consisted of integrated software for clinical use, and one was described as a primary care EHR. Six systems (6/17) were hospital EHRs. Among the system types were also the following: an "outpatient and inpatient EHR," a "centralized EHR with web-interface," and a "local EHR." One of the use cases described the system generally as an "ehr."

To further verify the clinical orientation of the use cases, we analyzed the professional groups involved in each of them. Although users were described, it was not clearly stated which professional groups were the intended users of SNOMED CT (eg, nursing informatics, medical informatics, or multi-professional users). Seven of the use cases (7/17) described the users in an exact way. Four cases (4/17) were

applied by physicians, 2 (2/17) by nurses, and 1 by physicians and nurses. Two of the cases (2/17) were applied by a multi-professional team of clinicians, clinical and medical informatics professionals, clinical domain experts, terminologists, and clinical coders. Two of the articles specifically concerned clinical coders. Two of the remaining cases were generally described as having been applied by "clinicians." Four of the cases (4/17) concerned researchers themselves, for whom specific clinical backgrounds were not reported. Contextual factors are presented in Multimedia Appendix 3.

SNOMED CT Use Purpose

All 17 articles described one or several use purposes for the terminology. Implementing SNOMED CT as a standard terminology in the EHR was typically grounded on clinical needs for standardizing patient information. Here, the SNOMED CT use purpose refers specifically to the primary use goal of the terminology as integrated or being implemented into an EHR. The most common category of use purpose (8/17) was SNOMED CT adopted as a common standard for EHRs. An additional 2 studies (2/17) described the goal of implementing SNOMED CT as a standard in a separate clinical application integrated in the EHR system. The use cases described communication and coordination needs, such as between hospital units or between inpatient and outpatient care, with the goal of promoting more reliable continuity of care and ultimately a higher quality of care. Accurate and timely diagnosis information with SNOMED CT deployment was reported as a crucial clinical need since it is major information in patient care. In 2 of the studies, SNOMED CT was implemented into the documentation of the problem list to increase the usefulness of the patient information and to organize the problem list content. Additionally, SNOMED CT was used to ensure effective data migration between systems.

The primary use purpose described in 2 articles was to retrieve or analyze patient data for clinical research. This enhanced data retrieval, analysis, and sharing for clinical research across multiple hospitals. Multisite data sharing and distributed analysis was supported by common terminology and by common data models. These 2 use cases utilized a medical annotation toolkit that included a web interface for extracting needed concepts. An additional 2 studies focused on data extraction, where SNOMED CT was used to classify and code patient data for research purposes. The use purpose in these 2 studies was building clinical pathways and patient selection criteria based on terminology coding. Natural language processing of clinical, pathology, and genomics data was used for further clinical research. Moreover, the use cases illustrated the challenges of data sharing between inpatient care and a virtual hospital visit.

Two of the original articles described, to a degree, the already established use of SNOMED CT with a focus on proving the merit of the terminology use in a clinical setting. The reasons for poor clinical coding of patient data after 2 decades of EHR use are manifold; two main reasons are lack of motivation and training. Support tools for the interoperable recording of diagnostic, treatment, and interventional patient information can be advanced, for example, with domain-specific development. One of the articles documented automated clinical coding as the driving purpose for SNOMED CT development. The development of computer-assisted coding may, through careful review and validation, improve the productivity of clinical coders. Different classification systems, such as the International Classification of Diseases–10, are typically linked and mapped to SNOMED CT for suitability in clinical use. An overview of the SNOMED CT use purposes is provided in Multimedia Appendix 3.

SNOMED CT Use Phase

The phase of use was identified in the literature with varying accuracy, which is why the research team discussed these categories during the analysis. Clinical use phase was documented in all 17 articles (100%), but in ambiguous ways. The most common category of use phases was SNOMED CT in development, which was documented in 6 articles (6/17). Development was described as an iterative process of analysis, validation, and standardization or building and mapping EHR-structured content that requires coordination and communication between stakeholders to improve the quality of care; as such, this was expected to be a process that could span several years.

SNOMED CT in use was identified as the use phase in 5 EHR-related use cases (5/17) and in implementation in 4 use cases (4/17). In the EHR-related use cases, SNOMED CT had been chosen as the base terminology system in the EHR or in a specific domain documented in the research to improve clinical information recording and coding, develop clinical pathways, and extract clinical data. The implementation cases addressed specifically improved the clinical recording of patient data by supporting clinicians' language and semantic selection with SNOMED CT or with a combination of SNOMED CT and other classification or terminology systems. In addition, 1 article documented a pilot use of SNOMED CT in EHR use cases, and 2 articles described the merits of SNOMED CT use with a more proven merit approach or through after-implementation evaluation. The pilot case evaluated cases of missing or mislabeled clinical data with, for example, nonstandard concepts or use of abbreviations. The evaluation of SNOMED CT use aimed to determine what had been achieved with fully integrated EHR services in patient care and to evaluate the impact of using SNOMED CT to record clinical meanings. As additional benefits, the secondary use purpose for using patient information was mentioned. An overview of the SNOMED CT use phase is provided in Multimedia Appendix 3.

Core Benefits of SNOMED CT

The research team identified and summarized core benefits of SNOMED CT as documented in the 17 use cases. The team categorized the benefits based on the previous literature (Multimedia Appendix 3). The core benefits were related to terminology use outcomes. The most common category was increased data quality, with 8 articles (8/17). Semantic-level core benefits were built on the scope and comprehensiveness of the terminology. In the use cases, SNOMED CT supported not only clinical meaning standardization but also the language of choice. For clinical use purposes, custom concept dictionaries or language-specific subsets were built for a chosen language.

Further benefits of implementing SNOMED CT were 2-fold: the parallel development of EHR technology and standardization. One UK use case documented evidence of increased interface usability and user satisfaction by clinicians. However, clinicians reported that adopting a new approach for data recording was a gradual process requiring time.

Four articles (4/17) documented the benefits category of enabling a consistent way of indexing, storing, retrieving, and aggregating clinical data. One use case concentrated on the benefits for data retrieval. Documented benefits pointed to the documentation of clinical events with richer detail. The productivity of coding was the main benefit categorized in 1 use case, increased quality of care in 2 use cases, and increased continuity of care in 1 use case. These benefits depended on the possibility of accessing more complete and coherent patient information, regardless of where it was recorded, to support safe patient care. Additionally, in 1 use case, the core benefit was successful implementation of a new EHR through harmonizing data structures. An overview of the SNOMED CT core benefits is provided in Multimedia Appendix 3.

Discussion

Summary of Findings

This systematic review identified 17 articles in which SNOMED CT was implemented and used in a clinical context in EHRs or related clinical applications. We aimed to confirm whether research has developed to allow for a shift of focus from previously published reviews that described potential use toward studies documenting plausible benefits of SNOMED CT. We present findings related to clinical use purposes, use phases, and core benefits of SNOMED CT over the last 5 years of ongoing efforts. These review categories are based on previous research [3,5,13] that provided a strong starting point for this analysis.

The use purpose for SNOMED terminology based on previous research (Table 1) was identified in all the articles reviewed. As we evaluated the use cases, these categories served our research material well. The most applied use purpose category was SNOMED CT as the planned standard for EHRs or other related applications. Other frequently applied use categories were the goals of using SNOMED for retrieving and analyzing patient data or implementing the terminology to advance the coding of patient data. Only 2 of the articles in the review entailed proof of merit of EHR implementation as the use purpose category. Based on these results, the initial observation was that there might be a level of interconnectedness between the use purpose and use phase. To prove this, data on the maturity of EHR solutions are needed to research the possible interconnectedness of EHR use and SNOMED CT. Moreover, it might be relevant to analyze relationships between use purpose and use phase. This requires testing the categories and their possible relationships with different data sets.

Regarding the use phase results, all the reviewed articles included descriptions of the SNOMED CT use phase, although details of related contextual factors, such as clinical environment, varied. This hampered the assessment of the overall picture of the use phase. Considering the results, this may be a feature of this specific material, being typical of EHR-related use cases of SNOMED CT. Thus, in the future, it may prove fruitful to pay particular attention to the descriptions of these types of SNOMED CT use cases. We propose to describe the phase of use in a more structured and contextual manner. This kind of accuracy would increase the scalability of the use-case results. Lee et al [3] have already highlighted that only a few SNOMED CT implementation cases are being published in the scientific literature. Through the systematic investigation of previous theoretical work [3], and with time, more comparable scientific publications on SNOMED CT use cases in a clinical context could be published.

Based on our review, there is still little research evidence on the benefits for clinical use of SNOMED CT in EHR-related use cases. We identified the following frequently reported categories of core benefits: improvement of data quality and enabling a consistent way of indexing, storing, retrieving, and aggregating clinical data. Closely related to these benefits were improvement in quality of care—with the goal of achieving better patient safety—and, based on better data quality, enabling better continuity of care. Additionally, the review found individual remarks on improving the productivity of coding though automation or through terminological support for clinical users. Such tools had potential to increase user satisfaction, although there was evidence for a need to involve clinicians from different domains in development. Evidence of practical advancement may motivate various clinical specialties to become more involved in SNOMED CT development work from early on.

Although previous research has categorized the possible benefits of SNOMED CT [3,5], the core benefits in our review were summarized by the research team. By doing this, we aimed to describe the specific benefits of the EHR-related use of SNOMED CT. The objective is to highlight that the categorization applied in this study (Table 1) requires further testing with different data to assess its validity. The set of categories applied by our research team was not proven to be comprehensive or exhaustive. Therefore, in future research, it could prove relevant to carefully evaluate SNOMED CT use cases from the perspective of typical benefits. Additionally, it could be important to research what kinds of disadvantages, risks, and bottlenecks can be detected. By evaluating different use cases, it might be possible to extract the general success factors of clinical SNOMED CT implementation. Overall, the evaluation of SNOMED CT implementation requires more attention. As an example, the European large-scale implementation of SNOMED CT, which is being funded by the European Commission, could at the same time advance evaluation studies or require systematic evaluation as a part of the funding process.

Our review revealed that the EHR system or related software were poorly described in most of the articles, and specific descriptions of the EHRs were scattered due to a lack of uniform structure for such descriptions. This is clearly an issue that would require more attention in future use-case descriptions, given the fact that SNOMED CT is designed to support the use of EHRs. To describe the capabilities and overall maturity of

the EHR system is core information in use case descriptions if the research result aims to benefit the clinical implementation process by avoiding previous obstacles and possible mistakes. We recognize that addressing potential mistakes does not mean that specific implementation experiences are universally generalizable, but that such implementation is required to be adapted to other clinical contexts. To promote the scalability of previous experiences, we suggest, for example, the application of the electronic medical record maturity model (EMRAM) in future use cases. EMRAM is a widely used tool developed by the Healthcare Information and Management Systems Society to measure the rate of adoption of EHR functions in health care settings. Its stages match the technological progress of the overall digitalization of the health care setting. Moreover, one possible starting point for future studies is to recognize specific use cases for software applications in clinical specialties, in which SNOMED CT would be valuable to accelerate and facilitate specific types of clinical implementations.

Limitations

This systematic review has limitations that could affect the plausibility of the results. The methods and results of our systematic review are transparently reported in detail to allow readers to assess the trustworthiness and applicability of our findings [16]. However, even though the review's methodological basis [16] is scientifically recommended, the varying levels of description in the research articles proved to be challenging; for example, the descriptions of background variables led to partial imprecision in the results. This especially affected the categories regarding the clinical context and type of EHR. The study's risk of bias was carefully considered during the research process, and no assumptions were made about missing or unclear information from the studies.

We aimed to avoid missing key studies and to minimize bias by conducting a thorough and comprehensive literature search. However, our search entailed only one database, PubMed. Nevertheless, according to recent literature, PubMed can be used as a principal search system [34]. PubMed is well suited for evidence synthesis in the form of systematic reviews since it meets all necessary performance requirements, such as the formulation of queries, the correct interpretation of queries by the system, and the reproducibility of searches. At the same time, it is impossible to be certain that a system that has proven to be successful in specific tests will not fail under different circumstances [34,35]. Moreover, not all SNOMED CT implementation projects are published in the scientific literature [3]. Thus, it is important to recognize that relevant information on clinical use cases may be found in different types of literature.

Conclusions

This literature review demonstrates that systematic reviews are relevant to the development of an understanding of SNOMED CT use and its possible benefits in further facilitating multi-professional, clinically driven implementations by summarizing essential findings based on evidence-based results. Clinical use cases are needed to promote the scalability of review results. To achieve the best out-of-use case reports, more emphasis should be placed on describing the contextual factors, such as the electronic health care system currently in use and the use of previous frameworks, to allow the comparability of results. Regarding future research, although other systematic reviews have addressed similar questions to ours, this review is necessary to shift the focus onto more clinically grounded implementation outcomes and benefits of the use of SNOMED CT. Generally, further research evidence is still needed to determine how exactly SNOMED CT benefits clinical care and patient information quality.

Acknowledgments

The authors acknowledge Finnish governmental research funding (TYH2019244) provided for this study.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Outline of the review protocol applied in this research.
[DOCX File , 20 KB - medinform_v11i1e43750_app1.docx]

Multimedia Appendix 2
Characteristics of the publications.
[DOCX File , 20 KB - medinform_v11i1e43750_app2.docx]

Multimedia Appendix 3
Overview of the results.
[DOCX File , 22 KB - medinform_v11i1e43750_app3.docx]

References

1. Overview of SNOMED CT. National Library of Medicine. URL: https://www.nlm.nih.gov/healthit/snomedct/snomed_overview.html [accessed 2023-01-27]
2. Lee D, Cornet R, Lau F, de Keizer N. A survey of SNOMED CT implementations. J Biomed Inform 2013 Feb;46(1):87-96 [FREE Full text] [doi: 10.1016/j.jbi.2012.09.006] [Medline: 23041717]
3. Lee D, de Keizer N, Lau F, Cornet R. Literature review of SNOMED CT use. J Am Med Inform Assoc 2014 Feb;21(e1):e11-e19 [FREE Full text] [doi: 10.1136/amiajnl-2013-001636] [Medline: 23828173]
4. Duarte J, Castro S, Santos M, Abelha A, Machado J. Improving quality of electronic health records with SNOMED. Procedia Technol 2014;16:1342-1350 [FREE Full text] [doi: 10.1016/j.protcy.2014.10.151]
5. Cornet R, de Keizer N. Forty years of SNOMED: a literature review. BMC Med Inform Decis Mak 2008 Oct 27;8 Suppl 1(Suppl 1):S2 [FREE Full text] [doi: 10.1186/1472-6947-8-S1-S2] [Medline: 19007439]
6. Vuokko R, Mäkelä-Bengs P, Hyppönen H, Lindqvist M, Doupi P. Impacts of structuring the electronic health record: Results of a systematic literature review from the perspective of secondary use of patient data. Int J Med Inform 2017 Jan;97:293-303. [doi: 10.1016/j.ijmedinf.2016.10.004] [Medline: 27919387]
7. Cangioli G, Chronaki C, Gooeg K, Højen A, Karlsson D, Jaulent M. Assessing SNOMED CT for large scale eHealth deployments in the EU. Community Research and Development Information Service. URL: https://assess-ct.eu/fileadmin/assess_ct/deliverables/final_submissions/assess_ct_ga_643818_d1.4.pdf [accessed 2023-01-27]
8. Chu L, Kannan V, Basit MA, Schaeflein DJ, Ortuzar AR, Glorioso JF, et al. SNOMED CT concept hierarchies for computable clinical phenotypes from electronic health record data: comparison of intensional versus extensional value. JMIR Med Inform 2019 Jan 16;7(1):e11487 [FREE Full text] [doi: 10.2196/11487] [Medline: 30664458]
9. Chang E, Mostafa J. The use of SNOMED CT, 2013-2020: a literature review. J Am Med Inform Assoc 2021 Aug 13;28(9):2017-2026 [FREE Full text] [doi: 10.1093/jamia/ocab084] [Medline: 34151978]
10. Upadhyay S, Opoku-Agyeman W. Factors that determine comprehensive categorical classification of EHR implementation levels. Health Serv Insights 2021;14:11786329211024788 [FREE Full text] [doi: 10.1177/11786329211024788] [Medline: 34188485]
11. Everson J, Rubin JC, Friedman CP. Reconsidering hospital EHR adoption at the dawn of HITECH: implications of the reported 9% adoption of a "basic" EHR. J Am Med Inform Assoc 2020 Aug 01;27(8):1198-1205 [FREE Full text] [doi: 10.1093/jamia/ocaa090] [Medline: 32585689]
12. HIMSS Adoption Model for Analytics Maturity (AMAM). Healthcare Information and Management System Society. URL: https://www.himssanalytics.org/amam [accessed 2023-01-27]
13. Gaudet-Blavignac C, Foufi V, Bjelogrlic M, Lovis C. Use of the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) for processing free text in health care: systematic scoping review. J Med Internet Res 2021 Jan 26;23(1):e24594 [FREE Full text] [doi: 10.2196/24594] [Medline: 33496673]
14. "Clinical" - English Meaning. Cambridge English Dictionary. URL: https://dictionary.cambridge.org/dictionary/english/clinical [accessed 2023-01-27]
15. Sittig DF, Wright A. What makes an EHR "open" or interoperable? J Am Med Inform Assoc 2015 Sep;22(5):1099-1101. [doi: 10.1093/jamia/ocv060] [Medline: 26078411]
16. Higgins J, Thomas J, Chandler J, Cumpston M, Li T, Page M. Cochrane Handbook for Systematic Reviews of Interventions version 6.2. Cochrane Training. URL: http://www.training.cochrane.org/handbook [accessed 2023-01-27]
17. Alahmar A, Crupi ME, Benlamri R. Ontological framework for standardizing and digitizing clinical pathways in healthcare information systems. Comput Methods Programs Biomed 2020 Nov;196:105559. [doi: 10.1016/j.cmpb.2020.105559] [Medline: 32531654]
18. Andersen S, Brandsborg C, Pape-Haugaard L. Use of Semantic Interoperability to Improve the Urgent Continuity of Care in Danish ERs. Stud Health Technol Inform 2021 May 27;281:203-207. [doi: 10.3233/SHTI210149] [Medline: 34042734]
19. Buendia O, Shankar S, Mahon H, Toal C, Menzies L, Ravichandran P, et al. Is it possible to implement a rare disease case-finding tool in primary care? A UK-based pilot study. Orphanet J Rare Dis 2022 Feb 16;17(1):54 [FREE Full text] [doi: 10.1186/s13023-022-02216-w] [Medline: 35172857]
20. Burrows EK, Razzaghi H, Utidjian L, Bailey LC. Standardizing clinical diagnoses: evaluating alternate terminology selection. AMIA Jt Summits Transl Sci Proc 2020;2020:71-79 [FREE Full text] [Medline: 32477625]
21. Gaudet-Blavignac C, Rudaz A, Lovis C. Building a shared, scalable, and sustainable source for the problem-oriented medical record: developmental study. JMIR Med Inform 2021 Oct 13;9(10):e29174 [FREE Full text] [doi: 10.2196/29174] [Medline: 34643542]
22. Hier DB, Pearson J. Two algorithms for the reorganisation of the problem list by organ system. BMJ Health Care Inform 2019 Dec;26(1):e100024 [FREE Full text] [doi: 10.1136/bmjhci-2019-100024] [Medline: 31848142]
23. Kraljevic Z, Searle T, Shek A, Roguski L, Noor K, Bean D, et al. Multi-domain clinical natural language processing with MedCAT: The Medical Concept Annotation Toolkit. Artif Intell Med 2021 Jul;117:102083. [doi: 10.1016/j.artmed.2021.102083] [Medline: 34127232]
24. Millares Martin P. Consultation analysis: use of free text versus coded text. Health Technol (Berl) 2021;11(2):349-357 [FREE Full text] [doi: 10.1007/s12553-020-00517-3] [Medline: 33520588]

25. Melman A, Maher CG, Needs C, Machado GC. Many people admitted to hospital with a provisional diagnosis of nonserious back pain are subsequently found to have serious pathology as the underlying cause. Clin Rheumatol 2022 Jun;41(6):1867-1871 [FREE Full text] [doi: 10.1007/s10067-022-06054-w] [Medline: 35015190]
26. Nguyen A, Truran D, Kemp M, Koopman B, Conlan D, O'Dwyer J, et al. Computer-assisted diagnostic coding: effectiveness of an NLP-based approach using SNOMED CT to ICD-10 mappings. AMIA Annu Symp Proc 2018;2018:807-816 [FREE Full text] [Medline: 30815123]
27. Pankhurst T, Evison F, Atia J, Gallier S, Coleman J, Ball S, et al. Introduction of systematized nomenclature of medicine-clinical terms coding into an electronic health record and evaluation of its impact: qualitative and quantitative study. JMIR Med Inform 2021 Nov 23;9(11):e29532 [FREE Full text] [doi: 10.2196/29532] [Medline: 34817387]
28. Pedrera M, Garcia N, Blanco A, Terriza A, Cruz J, Lopez E, et al. Use of EHRs in a tertiary hospital during COVID-19 pandemic: a multi-purpose approach based on standards. Stud Health Technol Inform 2021 May 27;281:28-32. [doi: 10.3233/SHTI210114] [Medline: 34042699]
29. Ryu B, Yoon E, Kim S, Lee S, Baek H, Yi S, et al. Transformation of pathology reports into the common data model with oncology module: use case for colon cancer. J Med Internet Res 2020 Dec 09;22(12):e18526 [FREE Full text] [doi: 10.2196/18526] [Medline: 33295294]
30. Sass J, Essenwanger A, Luijten S, Vom Felde Genannt Imbusch P, Thun S. Standardizing Germany's electronic disease management program for bronchial asthma. Stud Health Technol Inform 2019 Sep 03;267:81-85. [doi: 10.3233/SHTI190809] [Medline: 31483258]
31. Soguero-Ruiz C, Mora-Jiménez I, Ramos-López J, Quintanilla Fernández T, García-García A, Díez-Mazuela D, et al. An interoperable system toward cardiac risk stratification from ECG monitoring. Int J Environ Res Public Health 2018 Mar 01;15(3):428 [FREE Full text] [doi: 10.3390/ijerph15030428] [Medline: 29494497]
32. Wardle M, Spencer A. Implementation of SNOMED CT in an online clinical database. Future Healthc J 2017 Jun;4(2):126-130 [FREE Full text] [doi: 10.7861/futurehosp.4-2-126] [Medline: 31098449]
33. Willett D, Kannan V, Chu L, Buchanan J, Velasco F, Clark J, et al. SNOMED CT concept hierarchies for sharing definitions of clinical conditions using electronic health record data. Appl Clin Inform 2018 Jul;9(3):667-682 [FREE Full text] [doi: 10.1055/s-0038-1668090] [Medline: 30157499]
34. Cooper C, Booth A, Varley-Campbell J, Britten N, Garside R. Defining the process to literature searching in systematic reviews: a literature review of guidance and supporting studies. BMC Med Res Methodol 2018 Aug 14;18(1):85 [FREE Full text] [doi: 10.1186/s12874-018-0545-3] [Medline: 30107788]
35. Gusenbauer M, Haddaway NR. Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources. Res Synth Methods 2020 Mar;11(2):181-217 [FREE Full text] [doi: 10.1002/jrsm.1378] [Medline: 31614060]

Abbreviations

EHR: electronic health record
EMRAM: Electronic Medical Record Adoption Model
SNOMED CT: Systematized Nomenclature of Medicine–Clinical Terminology

Edited by C Lovis; submitted 23.10.22; peer-reviewed by J Hüsers, T Karen; comments to author 13.11.22; revised version received 05.12.22; accepted 22.12.22; published 06.02.23.

Please cite as:
Vuokko R, Vakkuri A, Palojoki S
Systematized Nomenclature of Medicine–Clinical Terminology (SNOMED CT) Clinical Use Cases in the Context of Electronic Health Record Systems: Systematic Literature Review
JMIR Med Inform 2023;11:e43750
URL: https://medinform.jmir.org/2023/1/e43750
doi:10.2196/43750
PMID:36745498

Original Paper

Structure of Health Information With Different Information Models: Evaluation Study With Competency Questions

Anna Rossander[1], MD; Daniel Karlsson[2], PhD

[1]Department of Applied IT, University of Gothenburg, Gothenburg, Sweden

[2]Swedish eHealth Agency, Stockholm, Sweden

Corresponding Author:
Anna Rossander, MD
Department of Applied IT
University of Gothenburg
Department of Applied Information Technology, Division of Informatics
Box 100
Gothenburg, 405 30
Sweden
Phone: 46 735989141
Email: anna.rossander@gu.se

Abstract

Background: There is a flora of health care information models but no consensus on which to use. This leads to poor information sharing and duplicate modelling work. The amount and type of differences between models has, to our knowledge, not been evaluated.

Objective: This work aims to explore how information structured with various information models differ in practice. Our hypothesis is that differences between information models are overestimated. This work will also assess the usability of competency questions as a method for evaluation of information models within health care.

Methods: In this study, 4 information standards, 2 standards for secondary use, and 2 electronic health record systems were included as material. Competency questions were developed for a random selection of recommendations from a clinical guideline. The information needed to answer the competency questions was modelled according to each included information model, and the results were analyzed. Differences in structure and terminology were quantified for each combination of standards.

Results: In this study, 36 competency questions were developed and answered. In general, similarities between the included information models were larger than the differences. The demarcation between information model and terminology was overall similar; on average, 45% of the included structures were identical between models. Choices of terminology differed within and between models; on average, 11% was usable in interaction with each other. The information models included in this study were able to represent most information required for answering the competency questions.

Conclusions: Different but same same; in practice, different information models structure much information in a similar fashion. To increase interoperability within and between systems, it is more important to move toward structuring information with any information model rather than finding or developing a perfect information model. Competency questions are a feasible way of evaluating how information models perform in practice.

(JMIR Med Inform 2023;11:e46477) doi:10.2196/46477

KEYWORDS

informatics; health care; information model; terminology; terminologies; interoperability; competency question; interoperable; competency; EHR; electronic health record; guideline; standard; recommendation; information system

Introduction

Background

Increased use of standards is often suggested as part of the solution to the problem of siloed and unusable information in electronic health records (EHRs) [1], but there is no consensus yet on what standards to use [2]. Instead, there is a flora of standards and the same information is structured with different standards in different settings [3-5]. There are different types of information standards. Some standards primarily aim to

structure information within systems (intraoperability), whereas some are geared toward sharing information (interoperability) [6], but often, both types of standards may be used in both settings. The standards differ between and within themselves regarding the "boundary problem" [7,8], that is, the demarcation between what information is structured with the information model and what is structured with terminology or values. The standards also differ regarding if terminology is stated or not and if so which terminology. Additionally, the terminologies are sometimes standards in themselves (eg, Logical Observation Identifiers Names and Codes [LOINC] or Systematized Nomenclature of Medicine Clinical Terms [SNOMED CT]) but sometimes system-specific or information model–specific value sets.

This combination of possibilities leads to a flora of informatics components, seemingly nonreusable between settings, as noted in previous works [2,9]. However, if the information models structure information in similar ways and there is some agreement on terminologies, perhaps a way forward would be to continue using different standards. Information exchange would be facilitated but not plug and play, as the content would be similar, and the workload of structuring health care information could be shared between users of different information models. Previous work has compared system configurations in relation to a single standard and showed that different system configurations could be unified [10,11]. Works comparing different standards have shown discrepancies in coverage and lack of alignment, primarily regarding terminologies [12,13]. Our hypothesis is that the differences between information models are overestimated. This work contributes by evaluating both the amount and type of differences between models and by providing and testing a method for comparing structure and terminology choices of different standards.

Aim

This work aims to explore if a possible solution to the challenge of sharing information and burden of modelling work within health care would be to continue using different information models. This work also aims to assess the usability of competency questions (CQs) for the evaluation of information models within health care.

Research Questions

The objective of this study was to answer the following 2 research questions:

1. How does the content of health care information differ between information models?
2. Is the method of CQs a feasible way of comparing content in information models?

Methods

Choice of CQs

There are quantitative methods to evaluate information structures in use today. For example, the CAMMS (Common Assessment Method for Standards and Specifications) [14] is an established guide in Europe for assessing a wide range of aspects with primarily a quantitative outcome. The aim of this work was, however, to examine *how* a sample of clinically relevant information is structured with different information models and not *how much* of the information the models could structure. To expose how information was structured, a method with qualitative results, that is, including structure and terminology of content, was needed. CQs have been used to evaluate ontologies for a long time [15]. In brief, the ontology is tested by selecting a relevant scenario and then posing questions to the ontology to see if and how the information needed to describe the scenario is structured within the ontology. CQs yield both quantitative and qualitative data. To our knowledge, CQs have not yet been used to evaluate the combination of information model and terminology within health care.

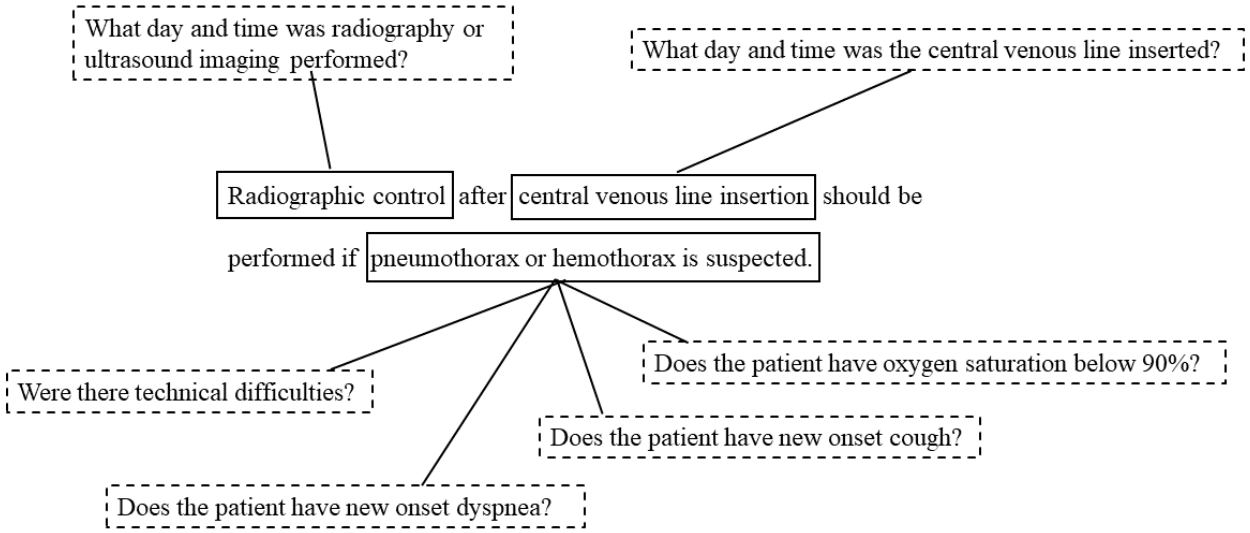
Development of CQs

Domain knowledge has been used as the basis for CQs previously by, for example, Cui [16]. Guidelines are an established textual source for domain knowledge within health care, and we thus chose to use recommendations in a guideline as scenario for developing the CQs. The use of information needed to follow best practice as a starting point ensured that the study examined the tested information models regarding clinically relevant information as opposed to theoretical possibilities or boundaries. Any topic within health care could have been used as a starting point for this work. Central venous lines (also called central catheters) are one of the many domains where structured documentation could support adherence to best practice and facilitate research to develop best practice. To prevent misinterpretations due to translation during the work, we chose to work with Swedish guidelines [17]. In the chosen guideline, there were 104 recommendations that covered preparation, insertion, care, and removal of central venous catheters. Examples from the guidelines are as follows:

1. The tip of the central venous line should be placed distally in the superior caval vein or the right atrium, and the location should be controlled at the time of insertion.
2. Bandages with polyurethane film should be replaced every 3-5 days during inpatient care.
3. At emergency insertion of a central venous line, the advantages of a central venous line should be weighed against the risk of hemorrhage.

The recommendations in the guideline were graded as beneficial, equivocal, or harmful. All recommendations graded as beneficial were placed in a random order and CQs were developed iteratively from the top. Formulating the CQs is a semantic task. The recommendation was read, and if needed, divided into sections, and then questions were developed to cover all the information mentioned in the recommendation. An example is shown in Figure 1.

Figure 1. An example of a recommendation (center) with developed competency questions (in dashed boxes).



Recommendations that contained information not documented in the patient record, for example, "All departments using central venous lines should have access to blood- and catheter tip–culture techniques" were considered out of scope in this study and omitted. When the recommendations were not specific enough to develop CQs, the associated text in the guideline was used to interpret and operationalize the recommendation. For example, for the recommendation "Radiographic control after central venous line insertion should be performed if pneumothorax or hemothorax is suspected," the text "Patients with pneumothorax who need treatment show new respiratory symptoms (dyspnea or cough) or oxygen saturation in blood lower than 90%" and "the risk increases with technical difficulties" was used to interpret the patients who had conditions indicative of pneumothorax or hemothorax.

The purpose of the study determines the number of CQs developed and used [18]. The focus of this work was on comparing how the different models structured clinically relevant information rather than the entire scope of each information model. During data collection, new CQs were iteratively developed and posed until further questions did not add additional types of clinical information. This is defined as the saturation point [19,20]. Despite this, a gap was discovered during data analysis regarding anatomical locations, which had not been covered by any of the initial CQs. Therefore, the next 2 recommendations in the randomized list containing information about the anatomical location were included as well. In total, 36 CQs based on 10 recommendations were developed. See Multimedia Appendix 1 for the list of included recommendations and developed CQs.

Materials

The information models tested and compared in this study were a purposive sample. In this study, the information models are the "participants" that were selected "based on the researchers' judgment about what potential participants will be most informative" [18]. The intention was to compare some typical models that were in use already and some models that were often recommended. The included information models differ in nature in several aspects, but they are all aimed at structuring data and thus impact interoperability. See Table 1 [21-32] for the included information models.

Table 1. Information models included in this study.

Information model	Description by information model provider
Information standards	
FHIR[a] [21]	FHIR is a standard for health care data exchange, published by HL7[b] [21]
openEHR [22]	openEHR is a nonprofit organization that publishes technical standards for an EHR platform along with domain-developed clinical models to define content [23]
HCIM[d] [24]	HCIMs are used to capture functional semantic (nontechnical) agreements for the standardization of information used in the care process [25]
IPS CDA[e] [26]	The goal of this project is to identify the required clinical data with associated vocabulary bindings and value sets for patient summary…and to build an international document and associated templates based on HL7 CDA R2…with value sets to support data elements within those templates [27]
Standards for secondary use	
OMOP[f] [28,29]	The OMOP Common Data Model allows for the systematic analysis of disparate observational databases. The concept behind this approach is to transform data contained within those databases into a common format (data model) as well as a common representation (terminologies, vocabularies, coding schemes) and then perform systematic analyses by using a library of standard analytic routines that have been written based on the common format [30]
SPOR[g] [31]	The purpose of SPOR is to, by means of integration with existing local operation planning systems, retrieve data from the perioperative process and thus offer a tool for local and national quality development (translation by authors) [32]
System-specific formats	
Electronic health record A	A health care information system used by approximately 70,000 health care staff
Electronic health record B	A health care information system developed and supplied by a global vendor

[a]FHIR: Fast Healthcare Interoperability Resources.

[b]HL7: Health Level 7.

[c]HCIM: Health and Care Information Model.

[d]IPS CDA: International Patient Summary Clinical Document Architecture.

[e]OMOP: Observational Medical Outcomes Partnership.

[f]SPOR: Svenskt Perioperativt Register.

Information about the information standards and standards for secondary use were sought on publicly available sources online. Fast Healthcare Interoperability Resources (FHIR) and openEHR have national and local profiles in addition to the internationally published standards available, for example, on Simplifier [33] and in national or local clinical knowledge manager repositories [22]. An initial survey of these resources did not show profiles directly focused on the application domain, and hence, these resources were not included. The information standards are in continuous development; the latest available version was used and cited (see individual references). Draft versions were included when there was no published version of a relevant component. The Health and Care Information Models (HCIMs) are a precursor for International Organization for Standardization (ISO) 13972 [34], which were not yet published as a standard when work began, and they were therefore used as an example of that standard. The International Patient Summary (IPS) is published as both Clinical Document Architecture (CDA) and FHIR. Since FHIR was included separately, the IPS CDA format was chosen.

Templates from 2 EHR systems were included. The material consisted of locally configured user interfaces of the systems and not an information model or database model; thus, the types of results differ between the 2 EHR systems and the other included models. Further, the 2 EHR systems offer a wide possibility for users to configure templates paired with limited reference information models, and thus, the results in this work provide examples of use in the selected EHR systems. The results might have been different if an application domain other than central venous lines had been used.

SPOR (Svenskt Perioperativt Register) and the 2 EHR systems often structure the information according to the specific situation where the templates are used, as opposed to the information standards and OMOP (Observational Medical Outcomes Partnership), which are intended to be general purpose. Thus, for SPOR and the EHR systems, we have included examples of data elements where generic elements do not exist. For example, SPOR had a data element for “kind of venous access,” with access devices in the value set, where the information standards often had a generic device type data element, which could hold any device type.

Answering the CQs

A table with the recommendations and corresponding CQs was developed. For each model, the authors together modelled the information needed to answer the CQs based on information available on the internet about the models. Both the structure of the information model, that is, what archetype/profile/entry

and element was used, and terminological content, that is, what code/codesystem/unit, was documented. As an illustration, the answers for the CQ "Does the patient have new onset dyspnea?" are displayed in Table 2. For details, please see Multimedia Appendix 2. Note that the CQ does not specify what "new" means, that is, in terms of hours or days, but to determine if a symptom is new by any definition, the time of onset is needed.

Table 2. Example results for "Does the patient have new onset dyspnea?" for selected information models.

Element and value	Value set
FHIR[a] condition resource	
condition.code = 267036007 |Dyspnea (finding)|	SNOMED CT[b] descendants of 404684003 |Clinical finding (finding)| (Example)
condition.onsetTime	ISO[c] 8601
IPS CDA[d] IPS problem entry	
hl7:value = 267036007 |Dyspnea (finding)|	SNOMED CT CORE Problem List Disorders (preferred)
hl7:effectiveTime	ISO 21090 → ISO 8601
OMOP[e] condition occurrence	
condition.concept.ID = 267036007 |Dyspnea (finding)| (no code for dyspnea in ICDo3)	SNOMED CT or ICDo3[f]
condition_start_date or condition_start_datetime	ISO 21090 → ISO 8601

[a]FHIR: Fast Healthcare Interoperability Resources.

[b]SNOMED CT: Systematized Nomenclature of Medicine Clinical Terms.

[c]ISO: International Organization for Standardization.

[d]IPS CDA: International Patient Summary Clinical Document Architecture.

[e]OMOP: Observational Medical Outcomes Partnership.

[f]ICDo3: International Classification of Diseases for Oncology Third Edition.

For some standards, the same information could be structured in several ways. For example, with FHIR, the information needed to answer, "What day and time was the central venous line inserted?" could be structured with both a Procedure Resource and a DeviceUseStatement. With openEHR, both Evaluation Medical Device and Action Procedure could be used. In these cases, all options were documented as results. The answers to the questions were influenced by the knowledge of the modelers answering them. The background knowledge that the authors have together was estimated to be comparable to that of a system implementer. One of the authors (DK) is, by training, a computer scientist and health informatician with experience in, for example, European Committee for Standardization and ISO standards and EHR system configuration as well as SNOMED CT. The other author (AR) is a medical doctor and health informatician with experience in structuring quality registers, SNOMED CT, and EHR system configuration. Since the authors performed the modelling, they were both researchers and participants at the same time. This gave extra insights and understanding of the work performed but also introduced a risk of bias. However, none of the authors have either any background or held any position with any of the above organizations that biased the results in any way.

Assessment of Coverage

The models were graded for content coverage by type of clinical information. Coverage was graded into "structured" if the information needed to answer the CQs for that type of information was structured. It was graded "partially structured" when only parts of the information were structured. "Not structured" was used when there was no structure and "missing" when the information was not present in the information model.

Assessment of Content Differences

Tables were developed for structure and coding. For each combination of models, the way or different ways the information was structured was evaluated. The number of possible ways for each model was used as the denominator and the number of ways that were similar enough to be used in interaction with the other model was used as the numerator, and the ratios of the 2 compared models were multiplied. For example, procedure type could be structured in only 1 way in FHIR but in 2 separate ways in SPOR (see Table S1 in Multimedia Appendix 2)—all 3 had a similar distribution of information between element and value, and this thus gave the following result: $1/1 \times 1/2 = 50\%$.

Another example is procedure status where HCIM had no status field but instead used a time stamp (see Table S2 in Multimedia Appendix 2). The other models, if anything, had a coded value for status, and the result for HCIM was thus 0 for all combinations for this type of information: $0/1 \times x/y = 0\%$.

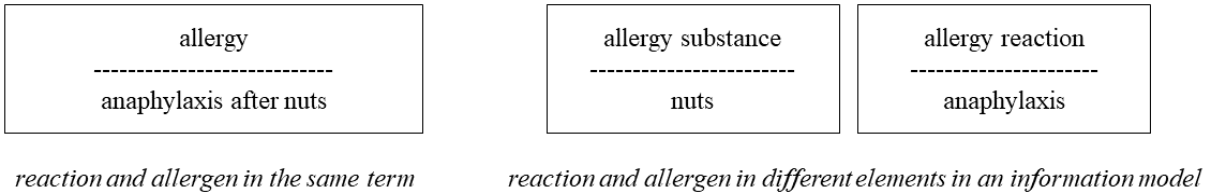
The value sets were assessed separately in the same fashion. Where a model had several possible terminologies, those that would have been used for the information needed to answer the CQs were used. The SNOMED CT Global Patient Set [35] and full SNOMED CT were considered usable in interaction, and LOINC terms that were in the same LOINC group were also considered sufficiently similar. For model-specific value sets, common in, for example, status elements, the included values were compared and if they were equivalent, this counted as

usable in interaction. All binding strengths have been assigned equal weight. When no value set was recommended, this was given the value 0.

Definitions

Information models, terminologies, and ontologies are all developed to structure information about things. In some subject areas, ontologies are in themselves sufficient to structure most information, whereas in health care, information models and terminologies or ontologies are usually used in conjunction. A division between information model and terminology is useful, as it makes the requirements on the terminology less complex. However, it leads to what is sometimes called the boundary problem, that is, the difficulty of deciding what information should be structured with the information model and what information should be structured with the terminology [7,8] (see example in Figure 2).

Figure 2. Examples of different ways of using information model and terminology in conjunction.



When not otherwise stated, we use the term information model for an information model, including its terminology bindings, that is, the terminologies or ontologies stated in it, when present. The term element is used for parts of the information model, sometimes also called attributes or headings. A value set is the stated terms or codes that are allowed for a certain element. Sometimes, a value set included entire terminology, for example, LOINC, International Classification of Diseases Tenth Revision (ICD-10), or SNOMED CT. Value set specifications may provide a binding strength to describe the flexibility with which members can be used while being compatible with the value set definition. For example, the FHIR framework provides 4 levels of binding strength: required (value set cannot be changed, eg, by extension), extensible (value set can be extended), preferred (the value set is recommended but not mandatory), and example (value set is an example only) [36]. For easier comparison of results, we have interpreted the binding strengths in the information models and described them using the FHIR definitions above.

Results

Research Question 1: How Does the Content of Health Care Information Differ Between Information Models?

The answers to the CQs included repeating types of information. The types of information are listed in Table 3, and the results below are presented per type. Note that the results depict the information needed to answer the CQs and not an overall evaluation of that information type.

Table 3. Types of information.

	Explanation	Example of competency question
Time and period	Point or period of time	What day and time was the central venous line inserted?
Procedures	Type, location, and status	Has the patient undergone radiography or ultrasound imaging?
Conditions	Type and status	Does the patient have renal impairment?
Causalities	Linking between elements or a specific causality element	Is the catheter occlusion considered due to thrombotisation?
Medications	Medicinal product	Has the patient received anticoagulant therapy?
Device types	Type or model of device used	What type of dressing was used?
Results of examinations	Type of examination and result	Does the patient have oxygen saturation below 90%?
Complex professional judgments	Assessments based on several discrete inputs and medical experience, often accompanied by motivation and degree of certainty	Is the patient's life at risk? What is the patient's need for central venous access?

Coverage

Results regarding coverage, that is, what information the included models had capacity to structure in a way that allowed answering of CQs, is provided per information model and type of information in Figure 3. With some exceptions, the differences in coverage were small between the included information models. The information standards and OMOP had the broadest coverage, providing structure for most types of information. SPOR could only structure information that was

requested when the registry was constructed. The EHRs could hold more information than the table implies, but some of the information was locked into structures, making it difficult to find or use it in other situations. For example, in EHR B, information about radiographic control after insertion of a central venous catheter could be found under the heading "use," where the options were "accepted for use," "may be used before radiographic control," "may not be used before radiographic control," and "other." Structures like this in EHR B were designed per instantiation and thus likely vary between settings.

Figure 3. Coverage per information model for competency questions. EHR: electronic health record; FHIR: Fast Healthcare Interoperability Resources; HCIM: Health and Care Information Model; IPS CDA: International Patient Summary Clinical Document Architecture; OMOP: Observational Medical Outcomes Partnership; SPOR: Svenskt Perioperativt Register.

	FHIR	open-EHR	HCIM	IPS CDA	OMOP	SPOR	EHR A	EHR B
Time and period	S	S	S	S	S	S	S	P
Procedure type	S	S	S	S	S	S	S	S
Procedure status	S	S	M	S	S	S	P	P
Procedure body location	P	P	P	P	N	P	P	P
Condition type	S	S	S	S	S	S	S	S
Condition status	S	S	S	S	S	S	S	S
Causality	S	S	N	N	N	N	P	P
Medications	S	S	S	S	S	M	S	S
Device types	S	S	S	S	S	P	P	P
Results of observations	S	S	S	S	S	S	S	S
Complex professional judgments	N	N	N	N	N	N	N	N

S: Structured
P: Partially structured
N: Not structured
M: Missing

Content Differences

Overall, the differences regarding the structure of information between the included information models were small. On average, 45% of the included structures were identical between models, that is, had the same demarcation between information model and terminology (Figure 4). The choice of terminology, however, showed a greater variation with, on average, only 11% overlap between models (Figure 5). Differences regarding structure were smaller than those regarding terminology (see Figure 6 for results per information type). Qualitative data on the content, that is, how information was structured and terminology was used, are presented as per the type of information below. The full result tables are provided in Multimedia Appendix 2. In the value set columns in the tables of Multimedia Appendix 2, all value sets listed within the information model are provided, although not all of them were relevant to the CQs. However, in our analysis, only the relevant value sets were considered.

Figure 4. Percentage of identical structures of information. EHR: electronic health record; FHIR: Fast Healthcare Interoperability Resources; HCIM: Health and Care Information Model; IPS CDA: International Patient Summary Clinical Document Architecture; OMOP: Observational Medical Outcomes Partnership; SPOR: Svenskt Perioperativt Register.

	Total average	Pairwise comparison average						
		openEHR	HCIM	IPS CDA	OMOP	SPOR	EHR A	EHR B
FHIR	50%	50%	50%	70%	70%	50%	40%	40%
openEHR	50%		50%	50%	50%	40%	30%	50%
HCIM	40%			50%	50%	40%	30%	40%
IPS CDA	50%				70%	50%	40%	40%
OMOP	50%					40%	40%	30%
SPOR	40%						20%	30%
EHR A	30%							30%
EHR B	40%							

Figure 5. Percentage of terminologies usable in interaction with each other. EHR: electronic health record; FHIR: Fast Healthcare Interoperability Resources; HCIM: Health and Care Information Model; IPS CDA: International Patient Summary Clinical Document Architecture; OMOP: Observational Medical Outcomes Partnership; SPOR: Svenskt Perioperativt Register.

	Total average	Pairwise comparison average						
		openEHR	HCIM	IPS CDA	OMOP	SPOR	EHR A	EHR B
FHIR	20%	10%	40%	50%	40%	0%	0%	10%
openEHR	10%		10%	10%	10%	0%	0%	0%
HCIM	20%			40%	20%	0%	0%	0%
IPS CDA	20%				40%	0%	0%	10%
OMOP	20%					0%	0%	10%
SPOR	0%						0%	0%
EHR A	0%							0%
EHR B	0%							

Figure 6. Average per information type for structure and terminologies.

	Identical structures	Terminologies usable in interaction
Time and period	N/A[a]	N/A
Procedure type	60%	10%
Procedure status	50%	0%
Procedure body location	20%	10%
Condition type	90%	30%
Condition status	70%	0%
Causality	0%	0%
Medications	40%	0%
Device types	40%	10%
Results of observations	40%	40%
Complex professional judgments	—[b]	—

[a] Not applicable.
[b] Not available.

Time and Period

Time is repeated in many different types of structures and thus not comparable in the same way as the other information types; hence, no table for time and period is provided in Multimedia Appendix 2. The information standards and OMOP used ISO 21090 [37] and ISO 8601 [38]. Since ISO 21090 is based on ISO 8601, they are equivalent in this setting. It was not possible to determine the exact format for the Swedish SPOR or the EHRs. Some modules in EHR B only handled time of documentation, as opposed to time of the actual event, procedure, or discovered condition. There were differences between the information models on how periods were represented. FHIR and IPS CDA used interval data types, while HCIM relied on having distinct data elements for start and end points of the period. openEHR had 3 different approaches. For action archetypes, periods could be deduced from the time difference between time-stamped events. For observation archetypes, time-related information was represented through the reference model, and for evaluation archetypes, distinct data elements were used for temporal information, similar to HCIMs.

Procedures

Information about procedures contain type of procedure (eg, insertion of central venous line), status of the procedure (eg, completed), and sometimes a location where the procedure was performed (eg, left subclavian vein). In FHIR, openEHR and OMOP procedures using a device could also be structured with the device as central information (for results regarding this, see "Device Types" below).

Procedure Type

All information standards and OMOP had a coded element within a dedicated procedure structure. SPOR and the 2 EHRs additionally had procedure-specific elements with a Boolean value. Of the information standards, none but HCIM strictly bound the procedure type to a terminology. HCIM had a binding strength of required and mandated the use of a code element in their ISO 21090–inspired CD datatype. The most commonly recommended terminologies in the included information models were SNOMED CT and the Swedish procedure classification KVÅ (Klassifikation av vårdåtgärder [39]; Swedish version and extension of Nordic Medico-Statistical Committee Classification of Surgical Procedures [40]) due to the Swedish context of some of the included information models. FHIR and IPS CDA used SNOMED CT, whereas HCIM and OMOP allowed several different value sets. The information models that used SNOMED CT pointed to different subsets. The Swedish registry SPOR and the 2 EHRs used KVÅ. They also, at times, used the term or code for the procedure as a question answered with a Boolean, for example, "C250 fluoroscopy during the procedure: yes/no."

Procedure Status

All models that had a stated status used a separate element for this. HCIM had no explicit representation of procedure status; instead, time could be both in the past and future, indicating performed or planned procedures. It was unclear how planned but not performed procedures could be discerned from performed procedures as time passes. The standards for secondary use only represented performed procedures, and this was sufficient for answering the CQs in this work. The EHRs had multiple structures. FHIR, openEHR, and IPS CDA used coded text with native value sets, which were not always one-to-one mappable between each other.

Procedure Body Location

The body location of a procedure can be represented either within the value for the procedure type (see above) or in a separate element. All models except OMOP had one or many ways to separately structure body location. FHIR, openEHR, HCIM, and SPOR also had additional elements for laterality or location qualifiers. In FHIR and openEHR, these were placed

in an extension and cluster, respectively. Many procedures have multiple possible locations, for example, regarding placing a central venous line–relevant location includes place of insertion (eg, left arm), the vessel in which the catheter is placed (eg, upper caval vein), and catheter tip location (eg, left atrium). None of the information models had a means to express the role of the body location in the procedure. All information standards, except openEHR, used SNOMED CT body structures. SPOR and the 2 EHRs used system-specific value sets.

Conditions

Condition Type

The demarcation between information model and terminology was identical for conditions in all the investigated models, except SPOR, which had an additional structure with separate Boolean elements for key conditions. Most of the compared information models used SNOMED CT as terminology, followed by ICD-10.

Condition Status

All information standards, OMOP, and EHR B, structured the status of the condition in a separate element. The information standards had at least 2 elements to capture both status (eg, present, resolved, absent) and certainty of the status (eg, unconfirmed, established, suspected). The difference between status of a condition and the certainty of the condition can lead to ambiguities; for example, in the FHIR Condition Resource, it was possible to have an active (clinicalStatus) and at the same time refuted (verificationStatus) condition. Two of the openEHR code sets and the code sets in HCIM contained codes from SNOMED CT but used different concepts, and the value sets were fully disjoint. All other codes were information model–specific. Most code sets had different granularity, that is, number of codes, making one-to-one mapping between them difficult.

Causality

Causality is a relation between entities where one is the cause of another, for example, that a deep vein thrombosis is a consequence of a central venous catheter. Of the included models, only openEHR had a separate element to document causality. openEHR also contained a LINK class, which would support this purpose, but there was no generic code set for the type of linking. In FHIR, there was a “dueTo” extension, which allowed linking conditions to their causes. The CDA standard had provisions for linking any CDA instance to any other instance, but this feature was not used for causality within the IPS CDA implementation guide. Both EHRs had structured lists for specific settings, for example, “reason for extraction” with local codes in the value set.

Medications

Only information regarding the type of medicinal product has been evaluated in this work. Information structures of timing, dosage, dose form, and substances were not included. The structure of information on medications varied depending on stage in the process of medication, that is, for example, prescribing, dispensing, administration, or consumption. There were 2 general patterns: one where there was a single coded element for the medicinal product and one where there was a complex structure of multiple elements, such as active ingredient, dose, and dose form. The information models that had terminology bindings pointed to multiple terminologies, except FHIR that stated SNOMED CT. EHR A used the Anatomic Therapeutic Chemical classification system and the Swedish national medicinal products terminology [41]; the configuration for EHR B was not finished at the time of data collection.

Device Types

Devices vary from short-time use artefacts as dressings to permanent implants as pacemakers. The results show how information about the type of devices was structured when the device was the central information. Several of the models had further elements for additional details, for example, batch number, size, or manufacturer. This was not included in this work. The information standards and OMOP had dedicated elements for devices with the name of the device in the value set, whereas SPOR and the EHRs used a terminology-bound element with a Boolean or a value set to further specify the device. FHIR, HCIM, IPS CDA, and OMOP all pointed to SNOMED CT as terminology. In general, it was also possible to document information about a device within the procedure where it was used as a distinct procedure type. This can be done with a term or concept where the device is included, for example, 1172566008 |Insertion of central venous catheter (procedure)|, but FHIR also permitted a separate element within the procedure class holding the device (in this example, 52124006 |Central venous catheter, device (physical object)|).

Results of Observations

Common observations are bedside measurements, assessment scales, and laboratory results. Information regarding observations is often a combination of a question, a result value, and a unit. Sometimes these entities were structured in separate elements, and sometimes, a part of the information was structured by terminology binding the element. For example, a measurement of the oxygen saturation could be structured into “Measurement = oxygen saturation, value = 98, and unit = %” as well as “oxygen saturation in percent = 98.” There were 2 distinct approaches to representing the results of the observations. FHIR, IPS CDA, OMOP, and EHR A rely on external terminologies to express the type of observation, whereas openEHR and HCIM develop specific information models to express the type of observation where the name of the element was bound to a terminology. SPOR and EHR B had specific elements in the information model for observations but without any terminology binding.

Complex Professional Judgments

None of the included models had a structured way to document complex judgments such as “How big problems can be expected if the central venous line is replaced?” or “Is the patient’s life at risk?”

Research Question 2: Are CQs a Feasible Way of Comparing Content in Information Models?

Development of CQs

In total, 36 CQs covering 10 recommendations were developed (Table 4). For 7 of the recommendations, the information in the recommendation was enough to develop the CQs—a task performed in a few minutes. For 2 recommendations, additional information from the guidelines was needed. One recommendation required information on what substances were included in "ADP (adenosine diphosphate) receptor antagonists" and "novel oral anticoagulants," which was not present in the guidelines.

Table 4. Information needed to develop the competency questions.

	Recommendations (n=10), n	Competency questions (n=36), n
Recommendation only	7	19
Recommendation and textual guideline	2	10
Additional information needed	1	7

As described in the Methods section, the initially assumed saturation point was revised during analysis of results, and additional CQs were developed for 2 recommendations.

Answering the CQs

The most effort in data gathering was spent on searching information about the information models and modelling. CQs covering information frequently documented in a structured way, for example, "Does the patient have renal impairment?" were relatively straightforward to answer with all the included models. For information that is rarely structured, for example, "Was a micro punction needle used?" or "What problems can be expected if the central venous line is replaced?" much time was spent on searching information about the different models to minimize risk that a possible solution was missed. The amount of work performed in modelling the information needed for the CQs is comparable to that performed in a real-life setting modelling clinical information. Time consumption thus varied widely both depending on complexity of the area and how well the chosen information model handled the area.

Assessment of Coverage and Content Differences

The results from the modelling work were complex, especially when information could be structured in several ways with the same information model or when terminology binding included multiple value sets. This was demanding to capture in a spreadsheet, but evaluation of tools was beyond the scope of this work.

Discussion

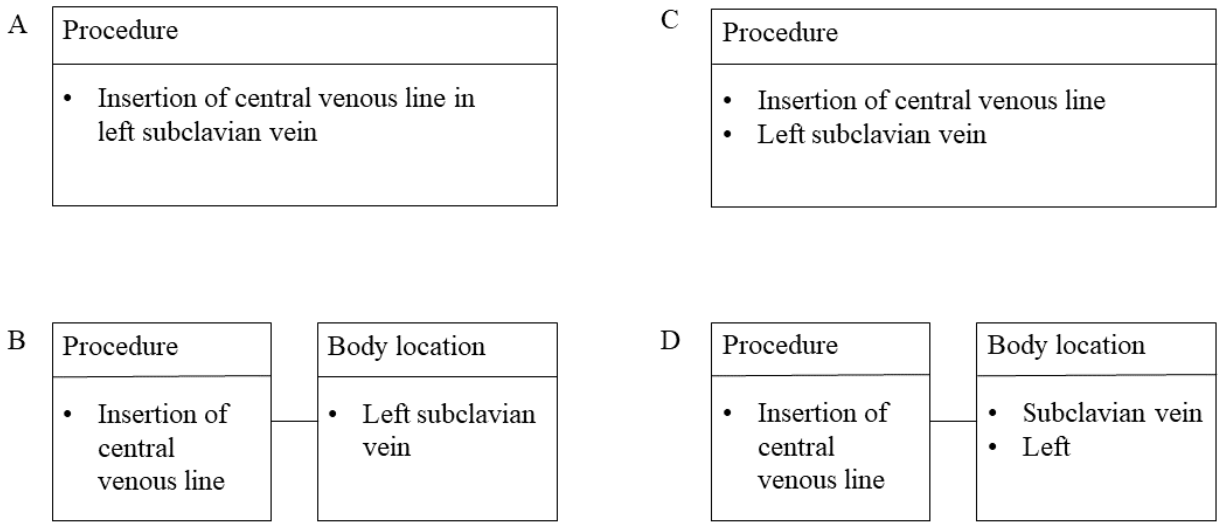
Research Question 1: How Does the Content of Health Care Information Differ Between Information Models?

When compared pairwise, the 8 included models had, on average, 45% identical structures and 11% terminologies that were sharable. Most overlaps regarding structure were present between information standards. Content that is not identical can still be similar, and our assessment is that the similarities were larger than the differences between the compared information models in general. The information models included in this study could represent most information required for answering the CQs.

Structure

Conditions and procedures have the highest overlap in structure. This information is thus readily sharable despite using different models if the used terminologies are the same or translatable. Representations of observations could be expected to be well-standardized due to its maturity but had only 40% overlap, mainly due to 2 different patterns of demarcation between information model and terminology. The same problem was present for medications and device types. Another common demarcation issue, present also for procedures and conditions, was that the EHRs and SPOR commonly used complex elements with a yes/no tick box, that is, with Boolean data type, whereas the information standards split the same information into several elements. Complex elements with tick boxes are tempting when developing structures for a specific use case but makes information sharing with structures developed for other use cases difficult. Conversion between these types of demarcation may be possible but includes risk of information loss and builds maintenance burden. In this material, the body locations for procedures and conditions could be represented either in a complex value for the procedure/condition itself (eg, insertion of central venous line in left subclavian vein or kidney failure) or separated in different ways (see examples in Figure 7). Similar issues can occur with other related information such as method.

Figure 7. Examples of using terminology or separate elements and classes. A. Using 1 element in 1 class instance. B. Using 2 elements in 2 class instances. C. Using 2 elements in 1 class instance. D. Using 3 elements in 2 class instances.



When sharing information structured with different demarcations, a compositional terminology for describing both the elements of the information model and the values in the model can be beneficial [42]. A compositional terminology allows for composition and decomposition of meaning, for example, splitting “kidney failure” into “organ failure” and “kidney” or vice versa. The information standards and OMOP all refer to SNOMED CT as a possible terminology for this type of information, and SNOMED CT logic representation may, in select cases, be used to transfer between different demarcations. For this to be possible, both the element name and the value must be terminology bound to concepts or postcoordinated expressions that are logically defined in relation to each other. Some use cases such as laterality are in that sense likely to be easier to coordinate, while others may introduce significant complexities regarding postcoordination or development of new concepts. Sometimes, different elements were mandatory in different models. In such cases, it might not be possible to share data even when the information is decomposable because obligatory information might be missing.

Terminologies

When the demarcation between information model and terminology is the same and the only difference between models is regarding terminology, information sharing possibilities depend on how easily those terminology-encoded values are converted into each other. All models used internal model-specific value sets for some elements. For example, in this material, the values for status for both procedures and conditions were different in all the included models, not only regarding terms but also by the number of values, making one-to-one mapping very difficult, not to say impossible, without information loss or distortion. Some values were, however, present in the code sets of all information models; for example, all models had a value to represent the status “the patient has this condition now” and that could thus be mapped between models. This confirms previous work, which showed that Apgar score representation had similar structures in HL7v3 DMIM (Health Level 7 version 3 Domain Message Information Model) and openEHR but were poorly bound to terminology [2] and that few value sets were aligned between the models when comparing openEHR and 3 HL7 formats for adverse sensitivity [12], although in the latter case, a joint openEHR-FHIR review has improved alignment [43].

The openEHR archetypes studied in this work were outliers compared to the other included information models in that few specific external terminologies were referred to. According to openEHR methodology, terminology binding is postponed to the templating phase, but while reviewing international templates, no additional terminology bindings were found. In this material, FHIR, HCIM, IPS CDA, and OMOP on the other hand often referred to external international terminologies, especially for larger value sets.

Where the models point to an existing terminology, there will be times when a suitable concept does not exist and therefore needs to be developed. In this material, there was, for example, no suitable concept within SNOMED CT to document the type of bandage used, despite SNOMED CT being the recommended terminology for several of the models. SNOMED CT provides the possibility to post coordinate concepts. However, postcoordination has drawbacks; for example, many health care information systems lack the capability to handle postcoordinated expressions, and postcoordinated expressions lack a human-readable term. Further, the concept model must permit the needed modelling, and the concepts needed for modelling must exist or be created [3]. Postcoordination has thus not been included as a possibility for value sets based on SNOMED CT.

Using the same terminology does not necessarily mean that the exact same code is used. For example, SNOMED CT contains 233527006 |Central venous cannula insertion (procedure)|, which has 16 more granular child concepts, and any of these could be used to document the insertion of a central venous

catheter where SNOMED CT is the recommended terminology. Lack of concepts, and to an extent, the lack of capabilities to postcoordinate have led to the development of national or implementation-specific extensions to many international terminologies, including SNOMED CT and ICD-10.

Internal Variability

The information standards aim to cover a wide range of information and offer complex structures to achieve this. They also sometimes have several different ways to structure the same information on varying levels of detail, leading to internal variability. This has been shown in evaluations of implementations of information models [13,44]. The standards for secondary use had a more rigid structure, only permitting 1 way to structure per type of information. The EHRs aim to capture all information and rely on free text to a higher degree than the other included types of models. Free text is, however, very hard to share unambiguously. Some information relevant in this work was structured very specifically in the EHRs, for example, "radiographic control before use of central venous catheter." Other types of radiography procedures were not examined but it is unlikely that all radiography examinations are structured like this, and this is thus an example of the same procedure being structured in multiple ways also in the EHRs.

In FHIR and openEHR, structured information could be added in extensions or slots. In our material, this was found for causality (FHIR), procedure body location (FHIR and openEHR), and medication detail (openEHR). These might be tempting if the other option is free text; however, additions like this risk add to complexity and internal variability. Where information can be structured in several ways, there is a risk that instantiated information is erroneous, for example, an "upper arm fracture in the leg." Having multiple elements to construct the meaning of a clinical statement increases the need for the sophisticated validation of information either during or after data entry to avoid mishaps.

Areas of Poor Coverage

Complex professional judgments were not possible to structure with any of the included models. Perhaps complex professional judgments are most easily documented as free text. Placing them in a terminology-bound element in an information model would facilitate identifying and sharing the information despite it being unstructured. openEHR and FHIR were the only included models that could structure causality, both by using extension and slots, thereby opening the potential for a higher degree of variability.

Research Question 2: Are CQs a Feasible Way of Comparing Content in Information Models?

The CQ method was a good way to probe deeper into information models from a clinically relevant perspective. The CQs revealed the types of information that were poorly structured or completely omitted—areas that are easily overseen when assessing the same information model from a theoretical perspective. One could argue that CQs leave the door ajar to bias from the evaluator, as opposed to a more formal method where the information models are described in the same format and then compared [45,46]. However, the information models relevant to compare are only available in different formalisms, often specific to the respective model, thereby restricting the use of such formal methods. Further, formal comparisons between models that differ in their demarcation between information model and terminology is not possible unless all elements are terminology bound to a machine-readable terminology, which they rarely are. Care should be taken when deciding how much effort to put into answering the CQs. Those covering information rarely structured are laborious to answer and perhaps more useful as a marker for where information is so complex that free text is the most suitable way to document it.

To cover all possible types of information, it might be necessary to push beyond the initially deemed saturation point. Perhaps, future work could use a 2-step saturation point by first developing CQs until no new types of information are uncovered and then modelling answers for the developed CQs, omitting those that duplicate already performed modelling work. This would avoid massive duplicative modelling.

Limitations

The CQs do not push the boundaries of what the information models can handle. For example, multiple body locations or status other than "performed" or "present" for procedures or conditions are not included nor are many of the intricacies about pharmaceutical information. There might be greater differences between models than this work has revealed. The information modelling done in this work is best effort but not best possible. The modelling was not discussed with additional parties external to this study; however, this might be in correspondence with results in a real-life setting, where the amount of effort is limited by existing resources, including access to domain and informatics experts.

Conclusions

Formal comparisons between information models show incompatibilities that are often merely theoretical [4,47], whereas practical work has shown that conversion between models for secondary use is doable [5,10,11]. This work shows that in practice, different information models structure much information in a similar fashion. To increase interoperability within and between systems, it is thus more important to move toward structuring information with any information model than finding or developing a single, perfect information model. When choosing an information model, one should consider that international standards have the best coverage and overlap between information models. They are also likely to be more widely adopted, decreasing the need for conversion before information exchange and have more users putting effort into developing them. As a final delimiter, assess the demarcation between information model and terminology and choose an information model that is similar to those of whom information is to be shared with. Put effort into decreasing internal variability and increasing terminology binding to external terminologies. The CQ method was successfully applied to the challenge of comparing health care information models. This method is a feasible way of evaluating how information models perform in practice, thereby adding valuable qualitative data on similarities and differences.

Acknowledgments

AR developed the competency questions. AR and DK performed modelling, analysis of results, and writing together. Thanks to Eva Blomqvist for contribution to initial ideas and anonymous reviewers for valuable comments.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Recommendations and corresponding competency questions.
[DOCX File , 17 KB - medinform_v11i1e46477_app1.docx]

Multimedia Appendix 2
Structure of content per type of information.
[DOCX File , 55 KB - medinform_v11i1e46477_app2.docx]

References

1. Evans RS. Electronic health records: then, now, and in the future. Yearb Med Inform 2018 Mar 06;25(S01):S48-S61. [doi: 10.15265/iys-2016-s006]
2. Cuggia M, Bayat S, Rossille D, Poulain P, Pladys P, Robert H, et al. Comparing the Apgar score representation in HL7 and openEHR formalisms. Stud Health Technol Inform 2009;150:250-254. [Medline: 19745308]
3. Rossander A, Lindsköld L, Ranerup A, Karlsson D. A state-of-the art review of SNOMED CT terminology binding and recommendations for practice and research. Methods Inf Med 2021 Dec;60(S 02):e76-e88 [FREE Full text] [doi: 10.1055/s-0041-1735167] [Medline: 34583415]
4. Goossen W, Goossen-Baremans A, van der Zel M. Detailed clinical models: A review. Healthc Inform Res 2010 Dec;16(4):201-214 [FREE Full text] [doi: 10.4258/hir.2010.16.4.201] [Medline: 21818440]
5. Pfaff ER, Champion J, Bradford RL, Clark M, Xu H, Fecho K, et al. Fast Healthcare Interoperability Resources (FHIR) as a meta model to integrate common data models: development of a tool and quantitative validation study. JMIR Med Inform 2019 Oct 16;7(4):e15199 [FREE Full text] [doi: 10.2196/15199] [Medline: 31621639]
6. Grieve G. Good exchange specifications: interoperability vs intraoperability. Health Intersect Pty Ltd. URL: http://www.healthintersections.com.au/?p=820 [accessed 2023-01-13]
7. Markwell D, Sato L, Cheetham E. Representing clinical information using SNOMED clinical terms with different structural information models. 2008 Presented at: Proceedings of the Third International Conference on Knowledge Representation in Medicine; May 31-June 2; Phoenix, Arizona, USA.
8. Rector A, Qamar R, Marley T. Binding ontologies and coding systems to electronic health records and messages. Journal of Applied Ontology 2009;1:51-69. [doi: 10.3233/ao-2009-0063]
9. Doyle-Lindrud S. The evolution of the electronic health record. CJON 2015 Apr 01;19(2):153-154. [doi: 10.1188/15.cjon.153-154]
10. Chen R, Klein GO, Sundvall E, Karlsson D, Åhlfeldt H. Archetype-based conversion of EHR content models: pilot experience with a regional EHR system. BMC Med Inform Decis Mak 2009 Jul 01;9:33 [FREE Full text] [doi: 10.1186/1472-6947-9-33] [Medline: 19570196]
11. Sundvall E, Terner A, Broberg H, Gillespie C. Configuration of input forms in EHR systems using spreadsheets, openEHR archetypes and templates. In: Studies in Health Technology and Informatics. Lyon, France: IOS Press; Aug 21, 2019:1781-1782.
12. Topaz M, Seger DL, Goss F, Lai K, Slight SP, Lau JJ, et al. Standard information models for representing adverse sensitivity information in clinical documents. Methods Inf Med 2018 Jan 08;55(02):151-157. [doi: 10.3414/me15-01-0081]
13. González-Ferrer A, Peleg M, Marcos M, Maldonado JA. Analysis of the process of representing clinical statements for decision-support applications: a comparison of openEHR archetypes and HL7 virtual medical record. J Med Syst 2016 Jul;40(7):163. [doi: 10.1007/s10916-016-0524-3] [Medline: 27209183]
14. Common Assessment Method for Standards and Specifications (CAMMS). European Commission. URL: https://joinup.ec.europa.eu/collection/common-assessment-method-standards-and-specifications-camss [accessed 2022-05-22]
15. Uschold M, Gruninger M. Ontologies: principles, methods and applications. The Knowledge Engineering Review 2009 Jul 07;11(2):93-136. [doi: 10.1017/s0269888900007797]
16. Cui H. Competency evaluation of plant character ontologies against domain literature. J Am Soc Inf Sci 2010 Jan 19;61(6):1144-1165. [doi: 10.1002/asi.21325]

17. Acosta S, Frykholm P, Granath A, Hammarskjöld F, Lindgren S, Lindwall R, et al. Riktlinjer för central venkateterisering. Svensk Förening för Anstesi och Intensivvård. 2018 Dec 03. URL: https://sfai.se/download-attachment/10389 [accessed 2021-07-01]
18. Moser A, Korstjens I. Series: Practical guidance to qualitative research. Part 3: Sampling, data collection and analysis. Eur J Gen Pract 2018 Dec;24(1):9-18 [FREE Full text] [doi: 10.1080/13814788.2017.1375091] [Medline: 29199486]
19. Guest G, Bunce A, Johnson L. How many interviews are enough? Field Methods 2016 Jul 21;18(1):59-82. [doi: 10.1177/1525822X05279903]
20. Locke K. Grounded Theory in Management Research. London: Sage Publications; 2001.
21. HL7. FHIR v4.3.0. URL: http://hl7.org/fhir/ [accessed 2022-07-08]
22. openEHR Clinical Knowledge Manager. URL: https://ckm.openehr.org/ckm/ [accessed 2022-07-08]
23. About us. openEHR. URL: https://www.openehr.org/about_us [accessed 2022-08-12]
24. HCIM Prerelease 2022-1(EN). URL: https://zibs.nl/wiki/HCIM_Release_2022(EN) [accessed 2022-07-08]
25. HCIM Mainpage. URL: https://zibs.nl/wiki/HCIM_Mainpage [accessed 2022-08-12]
26. International Patient Summary - Templates. URL: https://art-decor.org/art-decor/decor-templates--hl7ips-?section=templates [accessed 2022-07-08]
27. International Patient Summary - Project Information. URL: https://art-decor.org/art-decor/decor-project--hl7ips- [accessed 2022-08-12]
28. OMOP Common Data Model v5.4. URL: http://ohdsi.github.io/CommonDataModel/cdm54.html [accessed 2022-07-08]
29. The Book of OHDSI. URL: https://ohdsi.github.io/TheBookOfOhdsi/ [accessed 2022-07-08]
30. OMOP (Observational Medical Outcomes Partnership) common data model. Observational Health Data Sciences and Informatics. URL: https://www.ohdsi.org/data-standardization/the-common-data-model/ [accessed 2022-08-12]
31. Variabellistan. SPOR. URL: https://spor.se/spor-for-dig-som/vardgivare-tekniker/uppdatera-registret/ [accessed 2022-07-08]
32. Holmström B. SPOR Årsrapport 2020. URL: https://spor.se/wp-content/uploads/2021/09/Arsrapport-SPOR-2020_final.pdf [accessed 2022-08-12]
33. SIMPLIFIER.NET. The FHIR Collaboration Platform. URL: https://simplifier.net/ [accessed 2022-07-11]
34. Health informatics-detailed clinical models, characteristics and processes. ISO/TS 13972. 2015. URL: https://www.iso.org/standard/62416.html [accessed 2021-01-13]
35. Global patient set. Snomed International. URL: https://www.snomed.org/gps? [accessed 2023-04-28]
36. FHIR v4.3.0 Codesystem binding strength. HL7 FHIR. URL: http://hl7.org/fhir/R4B/codesystem-binding-strength.html [accessed 2022-10-31]
37. Health informatics - harmonized data types for information interchange. ISO 21090:2011. URL: https://www.iso.org/standard/35646.html [accessed 2022-07-06]
38. Date and time-representations for information interchange part 1: basic rules. ISO 8601-1. 2019. URL: https://www.iso.org/standard/70907.html [accessed 2022-07-06]
39. Klassifikation av vårdåtgärder (KVÅ). National Board of Health and Welfare. URL: https://www.socialstyrelsen.se/statistik-och-data/klassifikationer-och-koder/kva/ [accessed 2022-08-10]
40. Classification of surgical procedures. Nordic Health and Welfare Statistics. URL: https://nhwstat.org/publications/ncsp-classification-surgical-procedures [accessed 2022-08-10]
41. National substance register for medicinal products. Swedish Medical Products Agency. URL: https://www.lakemedelsverket.se/en/e-services-and-forms/substans-och-produktregister/national-substance-register-for-medicinal-products-nsl [accessed 2022-11-21]
42. Martínez-Costa C, Cornet R, Karlsson D, Schulz S, Kalra D. Semantic enrichment of clinical models towards semantic interoperability. The heart failure summary use case. J Am Med Inform Assoc 2015 Feb 10;22(3):565-576. [doi: 10.1093/jamia/ocu013] [Medline: 25670758]
43. Adverse reaction risk: the provenance. Archetypical. URL: https://omowizard.wordpress.com/2016/03/06/adverse-reaction-risk-the-provenance/ [accessed 2022-12-28]
44. Min L, Atalag K, Tian Q, Chen Y, Lu X. Verifying the feasibility of implementing semantic interoperability in different countries based on the openEHR approach: comparative study of acute coronary syndrome registries. JMIR Med Inform 2021 Oct 19;9(10):e31288 [FREE Full text] [doi: 10.2196/31288] [Medline: 34665150]
45. Martínez-Costa C, Schulz S. Validating EHR clinical models using ontology patterns. J Biomed Inform 2017 Dec;76:124-137 [FREE Full text] [doi: 10.1016/j.jbi.2017.11.001] [Medline: 29113934]
46. Martínez-Costa C, Menárguez-Tortosa M, Fernández-Breis JT, Maldonado JA. A model-driven approach for representing clinical archetypes for semantic web environments. J Biomed Inform 2009 Feb;42(1):150-164 [FREE Full text] [doi: 10.1016/j.jbi.2008.05.005] [Medline: 18590985]
47. Goossen W, Goossen-Baremans A. Bridging the HL7 Template - 13606 archetype gap with detailed clinical models. In: Studies in Health Technology and Informatics. Amsterdam, The Netherlands: IOS Press; 2010:932-936.

Abbreviations

CAMMS: Common Assessment Method for Standards and Specifications
CDA: Clinical Document Architecture
CQ: competency question
DMIM: Domain Message Information Model
EHR: electronic health record
FHIR: Fast Healthcare Interoperability Resources
HCIM: Health and Care Information Model
HL7: Health Level 7
ICD-10: International Classification of Diseases Tenth Revision
IPS: International Patient Summary
ISO: International Organization for Standardization
KVÅ: Klassifikation av vårdåtgärder
LOINC: Logical Observation Identifiers Names and Codes
OMOP: Observational Medical Outcomes Partnership
SNOMED CT: Systematized Nomenclature of Medicine Clinical Terms
SPOR: Svenskt Perioperativt Register

Edited by C Lovis; submitted 19.02.23; peer-reviewed by C Gaudet-Blavignac, J Ehrsam; comments to author 17.03.23; revised version received 11.05.23; accepted 03.06.23; published 31.07.23.

Please cite as:
Rossander A, Karlsson D
Structure of Health Information With Different Information Models: Evaluation Study With Competency Questions
JMIR Med Inform 2023;11:e46477
URL: https://medinform.jmir.org/2023/1/e46477
doi:10.2196/46477
PMID:37523221

Original Paper

The Impact of an Electronic Portal on Patient Encounters in Primary Care: Interrupted Time-Series Analysis

Karen Ferguson[1,2], MD, CCFP; Mark Fraser[2], MD, CCFP; Meltem Tuna[3,4], MSc, PhD; Charles Bruntz[5], MSc; Simone Dahrouge[1,6], MSc, PhD

[1]Department of Family Medicine, University of Ottawa, Ottawa, ON, Canada
[2]West Carleton Family Health Team, Carp, ON, Canada
[3]ICES, Ottawa, ON, Canada
[4]Ottawa Hospital Research Institute, Ottawa, ON, Canada
[5]Champlain Family Health Teams, Ottawa, ON, Canada
[6]Bruyère Research Institute, Ottawa, ON, Canada

Corresponding Author:
Karen Ferguson, MD, CCFP
Department of Family Medicine
University of Ottawa
600 Peter Morand Crescent #201
Ottawa, ON, K1G 5Z3
Canada
Phone: 1 613 562 5800 ext 2982
Email: karen@wcfht.ca

Abstract

Background: Electronic patient portals are online applications that allow patients access to their own health information, a form of asynchronous virtual care. The long-term impact of portals on the use of traditional primary care services is unclear, but it is an important question at this juncture, when portals are being incorporated into many primary care practices.

Objective: We sought to investigate how an electronic patient portal affected the use of traditional, synchronous primary care services over a much longer time period than any existing studies and to assess the impact of portal messaging on clinicians' workload.

Methods: We conducted a propensity-score–matched, open-cohort, interrupted time-series evaluation of a primary care portal from its implementation in 2010. We extracted information from the electronic medical record regarding age, sex, education, income, family health team enrollment, diagnoses at index date, and number of medications prescribed in the previous year. We also extracted the annual number of encounters for up to 8 years before and after the index date and provider time spent on secure messaging through the portal.

Results: A total of 7247 eligible portal patients and 7647 eligible potential controls were identified, with 3696 patients matched one to one. We found that portal registration was associated with an increase in the number of certain traditional encounters over the time period surrounding portal registration. Following the index year, there was a significant jump in annual number of visits to physicians in the portal arm (0.42 more visits/year vs control, $P<.001$) but not for visits to nurse practitioners and physician assistants. The annual number of calls to the practice triage nurses also showed a greater increase in the portal arm compared to the control arm after the index year (an additional 0.10 calls, $P=.006$). The average provider time spent on portal-related work was 5.7 minutes per patient per year.

Conclusions: We found that portal registration was associated with a subsequent increase in the number of some traditional encounters and an increase in clerical workload for providers. Portals have enormous potential to truly engage patients as partners in their own health care, but their impact on use of traditional health care services and clerical burden must also be considered when they are incorporated into primary care.

(JMIR Med Inform 2023;11:e43567) doi:10.2196/43567

KEYWORDS

electronic health records; health care utilization; patient portals; primary care; medical informatics; office visits; electronic; patient; online applications; virtual care; messaging; clinical; age; sex; education

Introduction

Electronic patient portals are online applications that allow patients access to their own health information, a form of asynchronous virtual care. There has been a great deal of recent interest in patient portals, accompanied by increasing technology adoption by both clinicians and patients [1-3]. The COVID-19 pandemic has also highlighted the importance of virtual care, an area already identified as a national health care priority [4]. Although portal features vary, the safe communication channels in portals may provide alternative ways for patients to obtain services traditionally provided in person, such as renewing prescriptions, sending and receiving secure messages, obtaining test results, and booking appointments [5]. A recent survey indicated that approximately 20% of Canadians had accessed some of their own medical information electronically, and that almost 80% were interested in doing so [6]. However, that survey did not specifically address portals or patient access to their medical information in primary care practice settings, and we are not aware of any studies examining Canadian portal adoption in primary care. Our understanding of the potential value of patient portals is nascent, with portals expected to contribute to more authentic collaboration between clinicians and patients.

The long-term impact of portals on traditional primary care services is unclear, but it is an important question at this juncture, when portals are being incorporated into primary care practices. Many studies reporting on the impact of portals on the use of traditional services evaluated systems that only provided options for web messaging or booking appointments [7-13]. All existing studies that investigated portals with more diverse features were conducted in medical networks, such as health maintenance organizations, where the portals provided access across sectors, including primary care, specialty care, and hospital care; these studies may not be relevant to portals incorporated into exclusively primary care practices. Past studies also reported inconsistent findings regarding the impact of portals on traditional health care use. Some studies demonstrated an increase in visits [14-16] or telephone calls [17]. Others demonstrated no change in visits [18], a reduction in visits [19], or a reduction in hospital readmissions [20]. All these studies also had limited time frames, examining only the period 12 to 30 months after portal registration.

To our knowledge, no long-term evaluation of the impact of a primary care patient portal on traditional health care use has been conducted to date. Providers have expressed interest in patient portals but also concerns regarding medicolegal risk and clerical workload [21]. Some have described an increased clerical burden associated with portals as part of the electronic health environment [12,22]. For instance, a qualitative study examining online patient access to their own health records found that providers felt that their workload had increased as a result [23], while another found that some providers anticipated fewer administrative requests for information when patients had access to their own health records [24]. One study found that online patient access to encounter notes did not significantly affect physician workload [25], although others have described high volumes of portal messages sent by patients [26]. However, no studies have actually tracked the provider time spent specifically on portal-related work. There have also not been any large studies of the impact of electronic patient portals in a Canadian setting. We sought to investigate how an electronic patient portal affects traditional, synchronous, primary care health care use over a much longer time period than any existing studies, and to assess the impact of portal messaging on clinicians' workload.

Methods

We conducted a propensity-score–matched, open-cohort, interrupted time-series (ITS) evaluation of a primary care portal from its implementation in 2010.

Setting and Study Participants

The practice was a semirural interprofessional clinic in southeastern Ontario, Canada, where 12 family physicians and other allied health providers provide comprehensive primary care under a single-payer model. Under this publicly funded model, physician compensation is primarily through capitation payments for rostered patients. The primary care patient portal initially offered access to laboratory results, the ability to enter vital signs such as blood pressure measurements, and the ability to view when certain screening maneuvers were due. Additional features were introduced over time, including the ability to receive secure messages (in 2012), send secure messages (in 2015), book appointments (in 2016), and renew prescriptions (in 2018.) All practice patients were invited to join the portal via email, posters, and telephone reminders and at in-person encounters.

We collected data for all practice patients except those seen exclusively for focused care (eg, obstetrical care). We retained data from all patients only for the period they were aged 18 years or older. Among patients who adopted the portal, we excluded those for whom we did not have at least one year of data prior to and following their portal registration (ie, index) date. For non–portal patients, we excluded those who did not have at least two consecutive years of data between 2009 and 2019.

Matching

We calculated propensity scores to estimate the probability of individuals registering for the portal using logistic regression [27,28]. Propensity scores were derived based on sex, age, whether the patient was rostered to the family health team, the presence of specific diagnoses on the index date, and the number of in-person and telephone encounters, as well as the number of unique medications prescribed in the 12 months prior to the index date. During the study time period, all appointments with medical doctors (MDs), nurse practitioners (NPs), and physician

assistants (PAs) were in person. Since education and income level were recorded for approximately a third of patients, these measures were not included in the propensity score matching. Control patients were entered into the equation for each year they were eligible (ie, for each year they had at least one year of data prior to and after the index date), with their corresponding profile for that year. July 1 of that year was considered the index date.

Variables, Data Sources, and Measurement

The study period was January 2002 through December 2019. We extracted electronic medical record information on patient age, sex, education, income, enrollment with the practice, and presence or absence of specific diagnoses on the index date. We also extracted the dates of in-person encounters with MDs, NPs, and PAs; dates of triage calls (TCs) to the practice triage nurses; and prescription dates and prescribed medications. Prescribed medications included only those that were identified as distinct medications using Anatomical Therapeutic Chemical codes. Diagnoses were defined based on diagnostic codes, using the earliest date when the diagnostic code was applied.

In order to study the clinician workload associated with the portal, two providers (KF and MF) time stamped their portal messages between February 20, 2020, and February 25, 2021. This allowed us to estimate the average provider time spent per message. We also collected the total number of portal messages sent by all providers to all portal patients between January 1, 2019, and December 31, 2019, in order to determine the average amount of time spent per patient on portal-related work.

Analyses

We described the profile of eligible patients prior to matching on their index date for portal patients on July 1 of the median year for which they were eligible to be matched for non–portal patients and again for the matched patients on their index date in both arms. The main study outcome was the frequency of in-person encounters with MDs, NPs, or PAs, as well as frequency of TCs. We used an ITS design to evaluate the impact of portal registration on use of these traditional health care services over time and compared use by portal users to their matched controls. We present the results in the usual ITS format, defining time relative to the index date with year 0 representing the 12 months preceding and including the index date and each time unit representing a 12-month interval. The ITS model includes time as a linear variable to model for an underlying linear time trend and the portal enrollment (ie, the intervention) as a dummy variable. Intervention and time interaction is also included in the model to identify the effect of the intervention on both arms (ie, portal and non–portal) over time.

We plotted the annual number of in-person encounters with MDs, in-person encounters with NPs or PAs, and TCs across time and overlaid the estimates derived from the ITS equations. During the study time period, all appointments with MDs, NPs, and PAs were in person. Because the year-0 results showed a spike in service use in both study groups, likely related to the attribution of the index date, we excluded that year from the ITS model. Also, although the spike in service use at year 1 in the portal arm may represent a transient change in behavior associated with the initial adoption, we also excluded this from the ITS to obtain a more reliable estimate of the impact of portal adoption over time, recognizing that this approach omits significant use; this should be considered in result interpretation.

We also depicted the number of visits per calendar year for patients who adopted the portal grouped by year of portal registration to demonstrate the pattern of changes in these visits over time for the intervention arm.

Ethics Approval

Ethics approval was received from the Bruyère Research Ethics Board (M16-20-012).

Results

Matching

Of the 14,894 patients who met the study criteria, 7247 (48.7%) were portal participants. Of these, 3696 were matched one to one with a control patient (Figure 1). The profile of all eligible patients before and after propensity matching is shown in Table 1. Before matching, portal users differed from non–portal users, but after matching, the mean propensity scores of the 2 groups and their index years, the prevalence of chronic conditions, sex, rostering status, and total visits and medications in the previous years showed good agreement. Income and education levels, which could not be included in the propensity score derivation because of poor data completeness, remained higher in the portal group.

We used a caliper of 0.2 for matching and limited the potential matching pool for each portal patient to non–portal patients with an index date that was within 1 year of the portal patient's index date. We identified all potential controls for each portal patient and assigned matches prioritizing first portal patients who had a unique match, then non–portal patients who had a unique portal match. We repeated this after each match to minimize loss of controls. When more than one match was possible, we attributed the control patient whose propensity score was closest to the portal patient's score. The balance of baseline covariates between the matched portal users and non–portal users was assessed using standardized differences, with values <0.1 representing negligible differences.

Figure 1. Study patient selection.

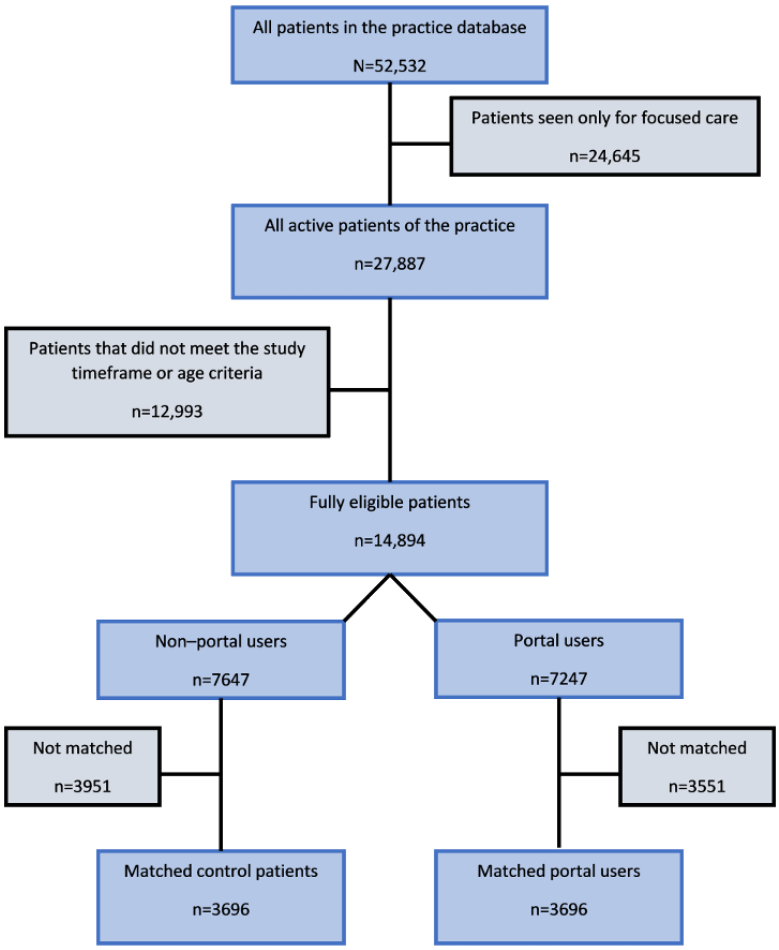


Table 1. Portal and control patients before and after matching.

Variable	Portal, index date (n=7247)	Non–portal, median index date (n=7647)	Total (n=14,894)	*P* value	Standard difference	Portal, index date (n=3696)	Control, index date (n=3696)	Total (n=7392)	*P* value	Standard difference
On index date										
Propensity score, mean (SD)	0.21 (0.10)	0.10 (0.08)	0.15 (0.11)	<.001	1.28	0.20 (0.10)	0.20 (0.10)	0.20 (0.10)	.82	0.01
Propensity score groups (participants), n (%)				<.001					.10	
0	291 (4)	2687 (35.1)	2978 (20)		0.85	738 (20)	740 (20)	1478 (20)		0
1	947 (13.1)	2032 (26.6)	2979 (20)		0.34	735 (19.9)	743 (20.1)	1478 (20)		0.01
2	1349 (18.6)	1631 (21.3)	2980 (20)		0.07	743 (20.1)	736 (19.9)	1479 (20)		0
3	2010 (27.7)	968 (12.7)	2978 (20)		0.38	740 (20)	742 (20.1)	1482 (20)		0
4	2650 (36.6)	329 (4.3)	2979 (20)		0.87	740 (20)	735 (19.9)	1475 (20)		0
Index year[a] (participants), n (%)				<.001					.99	
2010	1-5[b]	148-152[b]	153 (1)		0.2	1-5[b]	1-5[b]	1-5[b]		0
2011	1335 (18.4)	522 (6.8)	1857 (12.5)		0.35	654 (17.7)	654 (17.7)	1308 (17.7)		0
2012	1240 (17.1)	390 (5.1)	1630 (10.9)		0.39	579 (15.7)	596 (16.1)	1175 (15.9)		0.01
2013	948 (13.1)	349 (4.6)	1297 (8.7)		0.3	461 (12.5)	452 (12.2)	913 (12.4)		0.01
2014	531 (7.3)	3717 (48.6)	4248 (28.5)		1.04	296-300[b]	275-279[b]	576-580[b]		0.02
2015	664 (9.2)	838 (11)	1502 (10.1)		0.06	344 (9.3)	353 (9.6)	697 (9.4)		0.01
2016	970 (13.4)	670 (8.8)	1640 (11)		0.15	515 (13.9)	518 (14)	1033 (14)		0
2017	793 (10.9)	681 (8.9)	1474 (9.9)		0.07	464 (12.6)	455 (12.3)	919 (12.4)		0.01
2018	761-765[b]	328-332[b]	1093 (7.3)		0.24	378 (10.2)	388 (10.5)	766 (10.4)		0.01
Age at index date (years), mean (SD)	48.9 (14.9)	45.2 (19.3)	47.0 (17.4)	<.001	0.22	46.56 (15.17)	46.18 (15.93)	46.37 (15.56)	.29	0.02
Sex (participants), n (%)				<.001	0.22				.74	0.01
Female	4334 (59.8)	3748 (49)	8082 (54.3)			2104 (56.9)	2090 (56.5)	4194 (56.7)		
Male	2913 (40.2)	3899 (50.1)	6812 (45.7)			1592 (43.1)	1606 (43.5)	3198 (43.3)		
Rostered, n (%)	7162 (98.8)	7224 (94.5)	14,386 (96.6)	<.001	0.24	3665 (99.2)	3652 (98.8)	7317 (99)	.13	00.04
Coronary artery disease, n (%)	243 (3.4)	268 (3.5)	511 (3.4)	.61	0.01	68 (1.8)	79 (2.1)	147 (2)	.36	0.02
Congestive heart failure, n (%)	79 (1.1)	112 (1.5)	191 (1.3)	.04	0.03	24 (0.6)	31 (0.8)	55 (0.7)	.34	0.02
Chronic obstructive pulmonary disease, n (%)	128 (1.8)	235 (3.1)	363 (2.4)	<.001	0.09	43 (1.2)	54 (1.5)	97 (1.3)	.26	0.03
Diabetes mellitus, n (%)	440 (6.1)	482 (6.3)	922 (6.2)	.56	0.01	149 (4)	154 (4.2)	303 (4.1)	.77	0.01
Hypertension, n (%)	1343 (18.5)	1190 (15.6)	2533 (17)	<.001	0.08	495 (13.4)	474 (12.8)	969 (13.1)	.47	0.02
Income level (CAD $)[c,d] (participants), n (%)				<.001					<.001	
<40,000	233 (8.7)	432 (15.8)	665 (12.3)		0.22	121 (8.7)	150 (12.2)	271 (10.4)		0.11
40,000-60,000	357 (13.3)	484 (17.7)	841 (15.5)		0.12	173 (12.5)	218 (17.7)	391 (14.9)		0.15
60,000-100,000	846 (31.6)	836 (30.5)	1682 (31)		0.02	437 (31.5)	397 (32.3)	834 (31.9)		0.02
>100,000	1242 (46.4)	989 (36.1)	2231 (41.2)		0.21	657 (47.3)	465 (37.8)	1122 (42.9)		0.19
Education level[c] (participants), n (%)				<.001					<.001	

Variable	Portal, index date (n=7247)	Non–portal, median index date (n=7647)	Total (n=14,894)	*P* value	Standard difference	Portal, index date (n=3696)	Control, index date (n=3696)	Total (n=7392)	*P* value	Standard difference
High school or less	526 (21)	1037 (39.7)	1563 (30.5)		0.41	284 (21.7)	380 (32)	664 (26.6)		0.23
College	812 (32.4)	745 (28.5)	1557 (30.4)		0.08	397 (30.4)	381 (32.1)	778 (31.2)		0.04
University or more	1363 (50.5)	979 (35.5)	2342 (42.9)		0.31	626 (47.9)	427 (35.9)	1053 (42.2)		0.24
In the 12 months prior to index date										
Medical doctor visits[e], n (%)				<.001					.07	
0	1068 (14.7)	3328 (43.5)	4396 (29.5)		0.67	732 (19.8)	648 (17.5)	1380 (18.7)		0.06
1-2	3758 (51.9)	2548 (33.3)	6306 (42.3)		0.38	1969 (53.3)	2079 (56.3)	4048 (54.8)		0.06
3-5	1809 (25)	1251 (16.4)	3060 (20.5)		0.21	776 (21)	750 (20.3)	1526 (20.6)		0.02
6-10	532 (7.3)	437 (5.7)	969 (6.5)		0.07	189 (5.1)	187 (5.1)	376 (5.1)		0
>11	80 (1.1)	83 (1.1)	163 (1.1)		0	30 (0.8)	32 (0.9)	62 (0.8)		0.01
Nurse practitioner or physician assistant visits[e], n (%)				<.001					.34	
0	3959 (54.6)	5213 (68.2)	9172 (61.6)		0.28	2272 (61.5)	2297 (62.1)	4569 (61.8)		0.01
1-2	2655 (36.6)	2009 (26.3)	4664 (31.3)		0.22	1232 (33.3)	1234 (33.4)	2466 (33.4)		0
>3	633 (8.7)	425 (5.6)	1058 (7.1)		0.12	192 (5.2)	165 (4.5)	357 (4.8)		0.03
Calls to triage nurses[e], n (%)				<.001					.40	
0	4990 (68.9)	5886 (77)	10,876 (73)		0.18	2827 (76.5)	2871 (77.7)	5698 (77.1)		0.03
1-2	1898 (26.2)	1465 (19.2)	3363 (22.6)		0.17	759 (20.5)	728 (19.7)	1487 (20.1)		0.02
>3	359 (5)	296 (3.9)	655 (4.4)		0.05	110 (3)	97 (2.6)	207 (2.8)		0.02
Medications prescribed[e], n (%)				<.001					.78	
0	1705 (23.5)	3429 (44.8)	5134 (34.5)		0.46	1131 (30.6)	1132 (30.6)	2263 (30.6)		0
1-2	2522 (34.8)	2042 (26.7)	4564 (30.6)		0.18	1399 (37.9)	1439 (38.9)	2838 (38.4)		0.02
3-5	1928 (26.6)	1272 (16.6)	3200 (21.5)		0.24	817 (22.1)	792 (21.4)	1609 (21.8)		0.02
6-10	906 (12.5)	701 (9.2)	1607 (10.8)		0.11	288 (7.8)	281 (7.6)	569 (7.7)		0.01
>11	186 (2.6)	203 (2.7)	389 (2.6)		0.01	61 (1.7)	52 (1.4)	113 (1.5)	.82	0.02

[a]Index date for unmatched control patients: July 1 of the median year of their eligible time period.

[b]The value of n for 2010 was smaller than 6. The cells for 2010, 2014, and 2018 therefore do not have precise values due to ethics agreements.

[c]Income and education level were not available for all patients.

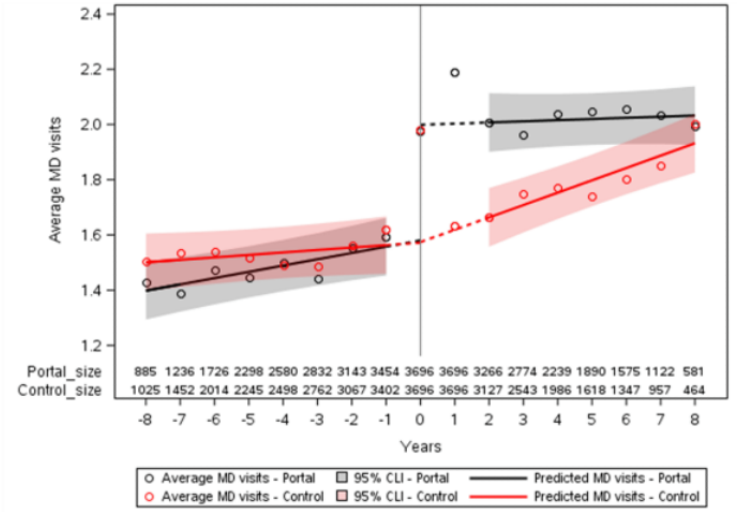
[d]CAD $1.00=US $0.75 on January 12, 2023.

[e]Number of medications prescribed and number of encounters refer to the 12-month period prior to the index date. The medication records were mapped to the Drug Product Database to assign Anatomical Therapeutic Chemical codes and schedules based on drug identification numbers. We excluded 12.5% of medication records, including records for medications classified as “over-the-counter” or “ethical” in the schedules from the Drug Product Database (4.9%) and medications reclassified manually as “over-the-counter,” “other,” or “N/A” (4.7%). Medications with no drug identification number were classified manually. Those which could not be attributed a drug identification number were also excluded (2.9%). Variable categorization for number of visits, telephone calls and medications was based on clinical judgement and the number of participants in each category.

Analyses

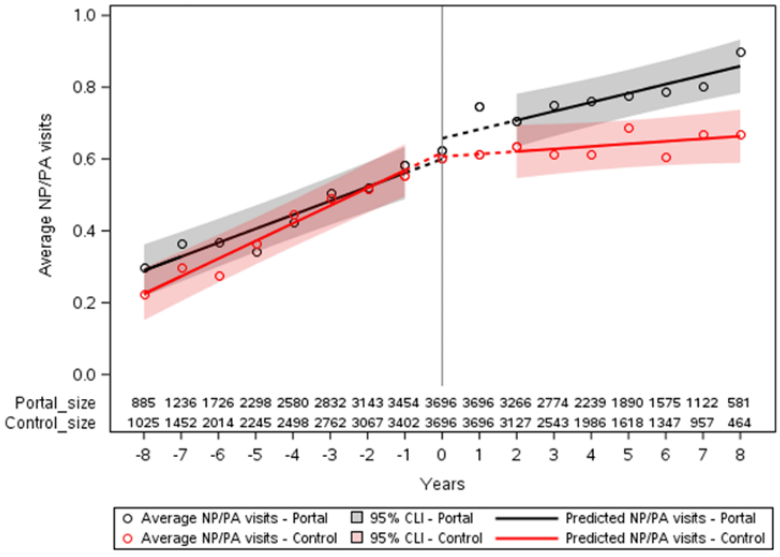
We plotted the number of visits in relation to the index year for portal and control patients, the estimated slopes for the years before and after the transition, and the shift in visits at the index date derived from the ITS equation (Figures 2-4). The information for the years prior to the index date for portal patients demonstrates their annual visits for the years prior to their registration on the portal. The index date for the control patients was assigned to be within 1 year of the portal’s patient index date in order to control for temporal factors such as health care use trends. The outputs of the ITS analyses are provided in Table 2. The intercepts and slopes prior to the index year were similar in the control and portal arms for MDs, NPs/PAs, and TCs (*P*>.05). After the index year, there was a significant jump in MD visits in the portal arm (0.42 more visits/year vs control, *P*<.001) but not for NP or PA visits. The TCs also showed a greater increase in visits in the portal arm compared to the control arm after the index year (0.102 more visits/year vs control, *P*=.006).

Figure 2. Interrupted time series for MD face-to-face visits for portal patients versus controls. The intercepts (*P*=.86) and slopes (*P*=.15) prior to the index year were similar in the control and portal arms. After the index year, there was no significant change in the number of MD visits in the control arm. However, in the portal arm, there was a significant jump in number of visits and a new intercept (0.42 more visits/year vs control, *P*<.001). The slope for MD visits increased after the index date in the control arm but became negative in the portal arm, representing an annual reduction of 0.054 visits per year for the portal arm compared to the control arm (*P*=.001). The two slopes would be expected to cross after 10 years. CLI: confidence limit interval; MD: medical doctor.



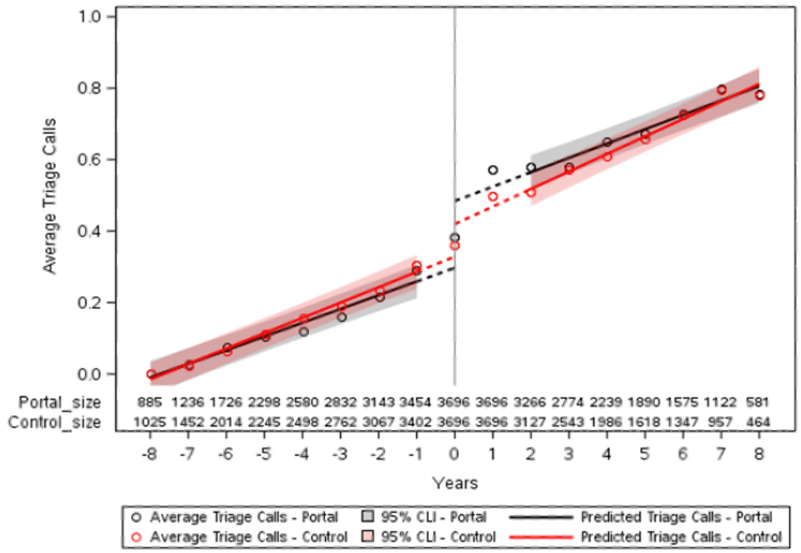
	Portal Patients	Control Patients	*P* value
Preintervention slope	0.023	0.0089	
Postintervention slope	-0.010	0.045	
Intervention absolute increase in annual # of visits	0.42	0.0022	<.001

Figure 3. Interrupted time series for nurse practitioner or physician assistant face-to-face visits for portal patients versus controls. The intercepts (*P*=.59) and slopes (*P*=.12) prior to the index year were similar in the control and portal arms. After the index year, there was not a significant change in the number of nurse practitioner or physician assistant visits in the portal arm compared to the control arm (*P*=.21). The slope flattened after the index date in the control arm, but it was relatively unchanged in the portal arm, demonstrating an annual increase of 0.028 visits per year in the portal arm compared to the control arm (*P*=.01). CLI: confidence limit interval; PA: physician assistant; NP; nurse practitioner.



	Portal Patients	Control Patients	*P* value
Preintervention slope	0.039	0.049	
Postintervention slope	0.036	0.0071	
Intervention absolute increase in annual # of visits	0.058	-0.011	.21

Figure 4. Interrupted time series for triage calls for portal patients versus controls. The intercepts (*P*=.10) and slopes (*P*=.26) prior to the index year were similar in the control and portal arms. The number of triage calls following the index year showed a higher value than anticipated based on the preindex slope in the control arm (0.062 more calls annually, *P*=.02), but a significantly greater jump after the index year in the portal arm (0.10 more calls annually, *P*=.006). The slopes for annual triage calls were similar in the pre- and postindex periods for both the control arm and portal arm. CLI: confidence limit interval.



	Portal Patients	Control Patients	*P* value
Preintervention slope	0.040	0.045	
Postintervention slope	0.050	0.055	
Intervention absolute increase in annual # of visits	0.16	0.062	.006

Table 2. Outputs of the interrupted time series. “Annual visits” indicates slope; “period” indicates the pre- or postindex period.

Variable	Medical doctor visits		Nurse practitioner or physician assistant visits		Triage calls	
	Estimate	*P* value	Estimate	*P* value	Estimate	*P* value
Intercept[a]	1.572	<.001	0.618	<.001	0.335	<.001
Annual visits (slope)[b]	0.009	.19	0.049	<.001	0.045	<.001
Period (before or after index)[c]	0.002	.97	-0.011	.77	0.062	.02
Annual visits × period[d]	0.036	.002	-0.042	<.001	0.010	.05
Portal[e]	0.008	.86	-0.018	.59	–0.033	.10
Portal × annual visits[f]	0.014	.15	-0.010	.12	–0.005	.26
Portal × period[g]	0.417	<.001	0.069	.21	0.102	.006
Portal × annual visits × period[h]	–0.054	.001	0.028	.01	–0.005	.46

[a]Control arm intercept.

[b]Pre–index date slope of annual visits for the control arm.

[c]Change in number of visits in year 2 post–index date relative to that anticipated from preindex slope for the control arm.

[d]Change in the slope of annual visits in the postindex period relative to the preindex period for the control arm.

[e]Difference between the portal arm and control arm in the intercept.

[f]Difference between the portal arm and control arm in the pre–index date slope of annual visits.

[g]Difference between the portal arm and control arm in the change in number of visits in year 2 post–index date relative to that anticipated from preindex slope.

[h]Difference between the portal arm and control arm in the change of the slope of annual visits in the postindex period relative to the preindex period.

We also plotted the visit rates for each year for patients having enrolled in the portal, grouped by year of portal registration (Figures 5-7).

The 2 physicians who time stamped 2061 portal messages spent an average of 3.83 minutes on each message. We also extracted the total number of portal messages sent by all providers between January 1 and December 31, 2019, and found that an average of 1.49 messages were sent to each portal patient in the practice. Thus, the average amount of provider time devoted to portal messages was estimated to be 5.7 minutes per portal patient per year.

Figure 5. Number of visits to medical doctors per calendar year for each patient group (registered on the portal in 2011-2012, 2013-2014, 2015-2016, and 2017-2018). To reduce noise, the number of visits represents the running average of that year, the previous year, and the following year. The MD visits showed a slight increase in the number of visits in the years immediately following portal registration, followed by an apparent drop in annual rate of visits.

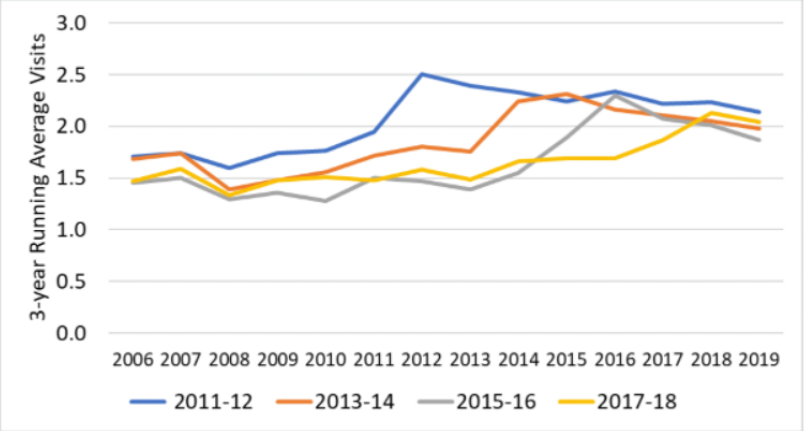


Figure 6. Number of visits to nurse practitioners or physician assistants for each patient group (registered on the portal in 2011-2012, 2013-2014, 2015-2016, and 2017-2018). To reduce noise, the number of visits represents the running average of that year, the previous year, and the following year. The NP and PA visits began in 2006 and show a rapid rise in the number of visits until 2010, then a considerable flattening of that slope afterwards with a potential small spike following the year of registration.

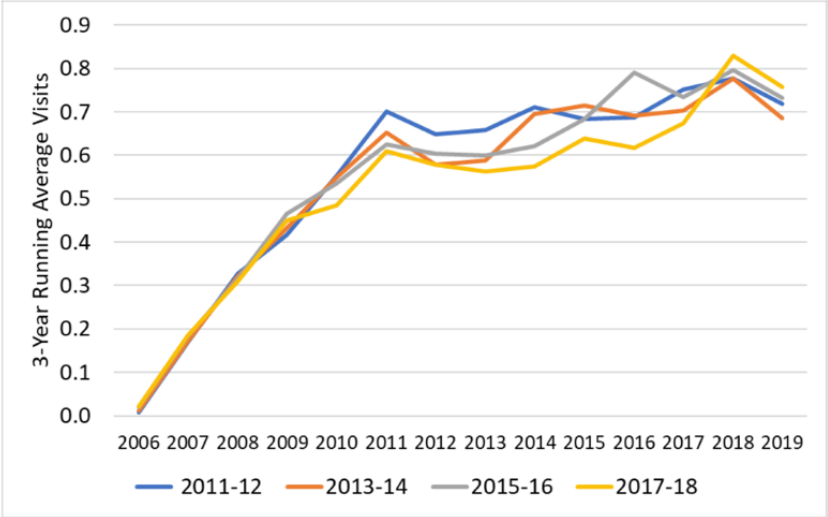
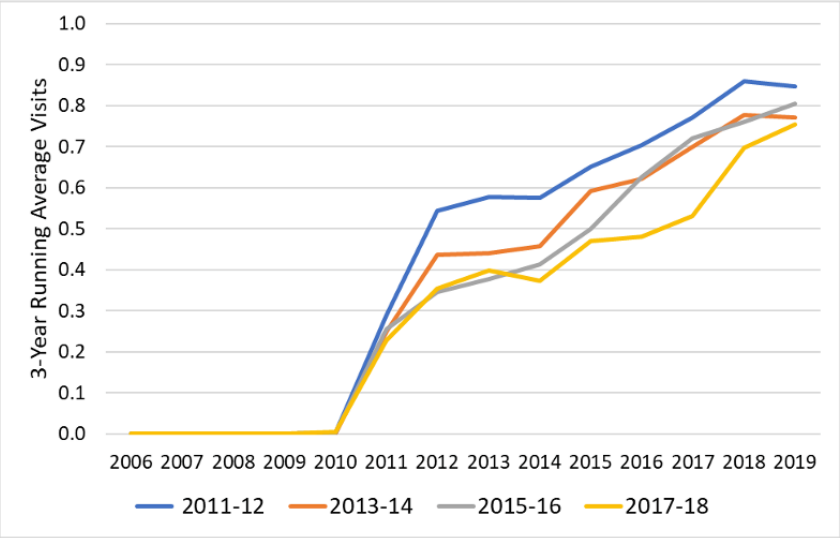


Figure 7. Number of triage calls per calendar year for each patient group (registered on the portal in 2011-2012, 2013-2014, 2015-2016, and 2017-2018). To reduce noise, the number of visits represents the running average of that year, the previous year, and the following year. The nurse triage calls were introduced in 2010 and show a consistent rise in frequency over time with a small increase in calls associated with the year of portal registration.



Discussion

Main Findings

Our findings suggest that portal registration is associated with an increase in service use, but that some reductions may be expected over subsequent years. Compared to matched controls, portal registration was associated with a significant initial increase in the number of in-person MD encounters and telephone calls, but a subsequent drop in the rate of MD visits and increase in NP visits over time. MDs spent an estimated 5.7 minutes per patient annually to respond to portal messages.

Limitations and Comparison With Prior Work

We believe that ours is the first study to examine the trend in encounters after portal registration over an extended time span and the first study to examine the impact of an exclusively primary care portal on traditional health care usage. It is possible that the observed increase in encounters was due to differences between the two groups that were not captured in the propensity matching. For instance, patients might have registered on the portal when they developed a new health concern, anticipating an increased requirement for health services. The reason for the gradual decrease in MD visits but increase in NP visits that took place after the initial jump in MD visits associated with portal registration is difficult to determine without further study. It is possible that patients initially presented to their own physician after sending them a portal message or viewing results, but the physician shared follow-up care with the nurse practitioner or physician assistant.

There may have been differences in areas such as electronic literacy or internet access that were not identified. It is also possible that the higher frequency of in-person encounters after portal registration was due to increased engagement by patients in their health. For instance, access to laboratory results may have generated questions from patients [29]. Increased awareness of being due for cancer screening or diabetes or blood pressure monitoring may have resulted in a higher number of encounters but improved quality of care or patient satisfaction. We did not examine these areas as they were beyond the scope of this study, but they would benefit from future research. While some past studies demonstrated improvements in certain health outcomes associated with electronic patient portals [30-33], only a few were based in primary care [31,34]. Several systematic reviews that evaluated a variety of portals in different practice settings suggested that portals or similar digital health services may result in improved patient satisfaction, but they did not demonstrate a meaningful impact on health outcomes, cost, or use [35-40].

We found that providers spent less than 6 minutes per year on clerical work for each patient registered on the portal. This is a small amount of time per patient but is significant when considering the context of an entire primary care practice. We note that the time-stamping of messages was performed during the COVID-19 pandemic, while the number of messages sent by all providers was collected prior to 2020. We consider that even if the COVID-19 pandemic resulted in an increased number of messages, the provider time per message would not have changed significantly. Therefore, our estimate of portal-related clerical work reflects prepandemic time requirements, and these may have increased since 2020 due to increased patient interest in asynchronous virtual care. This would also be an area for further study. Portals that do not allow incoming messages or any secure messaging would reduce or eliminate this time requirement but might also limit patient engagement and other potential benefits of the portal. Since the clerical burden associated with electronic environments in health care has been associated with professional burnout, [22,26] it is important to consider the provider time requirement associated with patient portals. The time and cost associated with incorporating a patient portal are currently not specifically addressed in either fee-for-service or capitated Canadian primary care funding models.

There are other limitations to this study. We examined the long-term impact of an electronic patient portal in a single

primary care practice, which may not be reflective of the impact in other primary care practices. However, portal adoption has not been widespread for long enough to allow study of the long-term impact of portals across multiple sites. Additionally, the impact of patient portals in other settings, such as hospitals, laboratories, or specialist practices, may be quite different. Further research is needed into electronic patient portals in different settings to determine their impact on various health outcomes.

Conclusions

Electronic patient portals are increasingly being adopted by providers and sought after by patients. We found that portal registration was associated with a subsequent increase in the number of some traditional encounters and a small increase in clerical workload for providers. Portals have enormous potential to truly engage patients as partners in their own health care, but their impact on use of traditional health care services and clerical burden must also be considered when they are incorporated into primary care.

Acknowledgments

This study was funded by a CAD $20,000 (US $15,000) research grant from the University of Ottawa Department of Family Medicine Program for Research and Innovation in Primary Care and Medical Education.

Conflicts of Interest

None declared.

References

1. Ryan B, Brown J, Terry A, Cejic S, Stewart M, Thind A. Implementing and using a patient portal: A qualitative exploration of patient and provider perspectives on engaging patients. J Innov Health Inform 2016 Jul 04;23(2):848 [FREE Full text] [doi: 10.14236/jhi.v23i2.848] [Medline: 27869582]
2. Stone JH. Communication between physicians and patients in the era of E-medicine. N Engl J Med 2007 Jun 14;356(24):2451-2454. [doi: 10.1056/NEJMp068198] [Medline: 17568026]
3. Gorfinkel I, Lexchin J. Enabling patient portals to access primary care medical records: maximizing collaboration in care between patients and providers. Healthc Policy 2019 May;14(4):21-27 [FREE Full text] [doi: 10.12927/hcpol.2019.25859] [Medline: 31322111]
4. Virtual Care: Recommendations for Scaling Up Virtual Medical Services. Canadian Medical Association. URL: https://www.cma.ca/sites/default/files/pdf/virtual-care/ReportoftheVirtualCareTaskForce.pdf [accessed 2023-01-12]
5. Archer N, Fevrier-Thomas U, Lokker C, McKibbon KA, Straus SE. Personal health records: a scoping review. J Am Med Inform Assoc 2011;18(4):515-522 [FREE Full text] [doi: 10.1136/amiajnl-2011-000105] [Medline: 21672914]
6. 2019 Canadian Digital Health Survey. Canada Health Infoway. URL: https://dataverse.scholarsportal.info/dataset.xhtml?persistentId=doi:10.5683/SP2/SPOAWK [accessed 2023-01-12]
7. Stamenova V, Agarwal P, Kelley L, Fujioka J, Nguyen M, Phung M, et al. Uptake and patient and provider communication modality preferences of virtual visits in primary care: a retrospective cohort study in Canada. BMJ Open 2020 Jul 06;10(7):e037064 [FREE Full text] [doi: 10.1136/bmjopen-2020-037064] [Medline: 32636284]
8. Bergmo TS, Kummervold PE, Gammon D, Dahl LB. Electronic patient-provider communication: will it offset office visits and telephone consultations in primary care? Int J Med Inform 2005 Sep;74(9):705-710. [doi: 10.1016/j.ijmedinf.2005.06.002] [Medline: 16095961]
9. Liederman EM, Lee JC, Baquero VH, Seites PG. Patient-physician web messaging. The impact on message volume and satisfaction. J Gen Intern Med 2005 Jan;20(1):52-57 [FREE Full text] [doi: 10.1111/j.1525-1497.2005.40009.x] [Medline: 15693928]
10. Zhou YY, Garrido T, Chin HL, Wiesenthal AM, Liang LL. Patient access to an electronic health record with secure messaging: impact on primary care utilization. Am J Manag Care 2007 Jul;13(7):418-424 [FREE Full text] [Medline: 17620037]
11. Shimada S, Hogan T, Rao S, Allison J, Quill A, Feng H, et al. Patient-provider secure messaging in VA: variations in adoption and association with urgent care utilization. Med Care 2013 Mar;51(3 Suppl 1):S21-S28. [doi: 10.1097/MLR.0b013e3182780917] [Medline: 23407007]
12. North F, Luhman KE, Mallmann EA, Mallmann TJ, Tulledge-Scheitel SM, North EJ, et al. A retrospective analysis of provider-to-patient secure messages: how much are they increasing, who is doing the work, and is the work happening after hours? JMIR Med Inform 2020 Jul 08;8(7):e16521 [FREE Full text] [doi: 10.2196/16521] [Medline: 32673238]
13. Meng D, Palen TE, Tsai J, McLeod M, Garrido T, Qian H. Association between secure patient-clinician email and clinical services utilisation in a US integrated health system: a retrospective cohort study. BMJ Open 2015 Nov 09;5(11):e009557 [FREE Full text] [doi: 10.1136/bmjopen-2015-009557] [Medline: 26553841]
14. Palen TE, Ross C, Powers JD, Xu S. Association of online patient access to clinicians and medical records with use of clinical services. JAMA 2012 Nov 21;308(19):2012-2019. [doi: 10.1001/jama.2012.14126] [Medline: 23168824]

15. Zhou Y, Leith W, Li H, Tom J. Personal health record use for children and health care utilization: propensity score-matched cohort analysis. J Am Med Inform Assoc 2015 Jul;22(4):748-754. [doi: 10.1093/jamia/ocu018] [Medline: 25656517]
16. Blok AC, Amante DJ, Hogan TP, Sadasivam RS, Shimada SL, Woods S, et al. Impact of patient access to online VA notes on healthcare utilization and clinician documentation: a retrospective cohort study. J Gen Intern Med 2021 Mar;36(3):592-599 [FREE Full text] [doi: 10.1007/s11606-020-06304-0] [Medline: 33443693]
17. Dexter EN, Fields S, Rdesinski RE, Sachdeva B, Yamashita D, Marino M. Patient-provider communication: does electronic messaging reduce incoming telephone calls? J Am Board Fam Med 2016;29(5):613-619 [FREE Full text] [doi: 10.3122/jabfm.2016.05.150371] [Medline: 27613794]
18. Leveille S, Mejilla R, Ngo L, Fossa A, Elmore J, Darer J, et al. Do patients who access clinical information on patient internet portals have more primary care visits? Med Care 2016 Jan;54(1):17-23. [doi: 10.1097/MLR.0000000000000442] [Medline: 26565525]
19. Zhong X, Liang M, Sanchez R, Yu M, Budd PR, Sprague JL, et al. On the effect of electronic patient portal on primary care utilization and appointment adherence. BMC Med Inform Decis Mak 2018 Oct 16;18(1):84 [FREE Full text] [doi: 10.1186/s12911-018-0669-8] [Medline: 30326876]
20. Martínez Nicolás I, Lê Cook B, Flores M, Del Olmo Rodriguez M, Hernández Rodríguez C, Llamas Sillero P, et al. The impact of a comprehensive electronic patient portal on the health service use: an interrupted time-series analysis. Eur J Public Health 2019 Jun 01;29(3):413-418. [doi: 10.1093/eurpub/cky257] [Medline: 30544169]
21. Mehta S, Jamieson T, Ackery A. Helping clinicians and patients navigate electronic patient portals: ethical and legal principles. CMAJ 2019 Oct 07;191(40):E1100-E1104 [FREE Full text] [doi: 10.1503/cmaj.190413] [Medline: 31591096]
22. Shanafelt TD, Dyrbye LN, Sinsky C, Hasan O, Satele D, Sloan J, et al. Relationship between clerical burden and characteristics of the electronic environment with physician burnout and professional satisfaction. Mayo Clin Proc 2016 Jul;91(7):836-848. [doi: 10.1016/j.mayocp.2016.05.007] [Medline: 27313121]
23. Turner A, Morris R, McDonagh L, Hamilton F, Blake S, Farr M, et al. Unintended consequences of patient online access to health records: a qualitative study in UK primary care. Br J Gen Pract 2023 Jan;73(726):e67-e74 [FREE Full text] [doi: 10.3399/BJGP.2021.0720] [Medline: 36316163]
24. Louch G, Albutt A, Smyth K, O'Hara JK. What do primary care staff think about patients accessing electronic health records? A focus group study. BMC Health Serv Res 2022 Apr 29;22(1):581 [FREE Full text] [doi: 10.1186/s12913-022-07954-y] [Medline: 35488233]
25. Delbanco T, Walker J, Bell SK, Darer JD, Elmore JG, Farag N, et al. Inviting patients to read their doctors' notes: a quasi-experimental study and a look ahead. Ann Intern Med 2012 Oct 02;157(7):461-470 [FREE Full text] [doi: 10.7326/0003-4819-157-7-201210020-00002] [Medline: 23027317]
26. Chavez A, Bracamonte J, Kresin M, Yardley M, Grover M. High volume portal usage impacts practice resources. J Am Board Fam Med 2020;33(3):452-455 [FREE Full text] [doi: 10.3122/jabfm.2020.03.190401] [Medline: 32430378]
27. Austin PC. Comparing paired vs non-paired statistical methods of analyses when making inferences about absolute risk reductions in propensity-score matched samples. Stat Med 2011 May 20;30(11):1292-1301 [FREE Full text] [doi: 10.1002/sim.4200] [Medline: 21337595]
28. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behav Res 2011 May;46(3):399-424 [FREE Full text] [doi: 10.1080/00273171.2011.568786] [Medline: 21818162]
29. Pillemer F, Price RA, Paone S, Martich GD, Albert S, Haidari L, et al. Direct release of test results to patients increases patient engagement and utilization of care. PLoS One 2016;11(6):e0154743 [FREE Full text] [doi: 10.1371/journal.pone.0154743] [Medline: 27337092]
30. Green BB, Cook AJ, Ralston JD, Fishman PA, Catz SL, Carlson J, et al. Effectiveness of home blood pressure monitoring, Web communication, and pharmacist care on hypertension control: a randomized controlled trial. JAMA 2008 Jun 25;299(24):2857-2867 [FREE Full text] [doi: 10.1001/jama.299.24.2857] [Medline: 18577730]
31. Nagykaldi Z, Aspy CB, Chou A, Mold JW. Impact of a Wellness Portal on the delivery of patient-centered preventive care. J Am Board Fam Med 2012;25(2):158-167 [FREE Full text] [doi: 10.3122/jabfm.2012.02.110130] [Medline: 22403196]
32. Ralston J, Hirsch I, Hoath J, Mullen M, Cheadle A, Goldberg H. Web-based collaborative care for type 2 diabetes: a pilot randomized trial. Diabetes Care 2009 Feb;32(2):234-239 [FREE Full text] [doi: 10.2337/dc08-1220] [Medline: 19017773]
33. Wright A, Poon EG, Wald J, Feblowitz J, Pang JE, Schnipper JL, et al. Randomized controlled trial of health maintenance reminders provided directly to patients through an electronic PHR. J Gen Intern Med 2012 Jan;27(1):85-92 [FREE Full text] [doi: 10.1007/s11606-011-1859-6] [Medline: 21904945]
34. Tom JO, Chen C, Zhou YY. Personal health record use and association with immunizations and well-child care visits recommendations. J Pediatr 2014 Jan;164(1):112-117. [doi: 10.1016/j.jpeds.2013.08.046] [Medline: 24120019]
35. Ammenwerth E, Schnell-Inderst P, Hoerbst A. The impact of electronic patient portals on patient care: a systematic review of controlled trials. J Med Internet Res 2012 Nov 26;14(6):e162 [FREE Full text] [doi: 10.2196/jmir.2238] [Medline: 23183044]

36. Goldzweig CL, Orshansky G, Paige NM, Towfigh AA, Haggstrom DA, Miake-Lye I, et al. Electronic patient portals: evidence on health outcomes, satisfaction, efficiency, and attitudes: a systematic review. Ann Intern Med 2013 Nov 19;159(10):677-687. [doi: 10.7326/0003-4819-159-10-201311190-00006] [Medline: 24247673]
37. Kruse CS, Bolton K, Freriks G. The effect of patient portals on quality outcomes and its implications to meaningful use: a systematic review. J Med Internet Res 2015 Feb 10;17(2):e44 [FREE Full text] [doi: 10.2196/jmir.3171] [Medline: 25669240]
38. Zanaboni P, Fagerlund AJ. Patients' use and experiences with e-consultation and other digital health services with their general practitioner in Norway: results from an online survey. BMJ Open 2020 Jun 17;10(6):e034773 [FREE Full text] [doi: 10.1136/bmjopen-2019-034773] [Medline: 32554721]
39. Mold F, de Lusignan S, Sheikh A, Majeed A, Wyatt JC, Quinn T, et al. Patients' online access to their electronic health records and linked online services: a systematic review in primary care. Br J Gen Pract 2015 Mar 02;65(632):e141-e151. [doi: 10.3399/bjgp15x683941]
40. Neves AL, Freise L, Laranjo L, Carter AW, Darzi A, Mayer E. Impact of providing patients access to electronic health records on quality and safety of care: a systematic review and meta-analysis. BMJ Qual Saf 2020 Dec;29(12):1019-1032 [FREE Full text] [doi: 10.1136/bmjqs-2019-010581] [Medline: 32532814]

Abbreviations

ITS: interrupted time-series
MD: medical doctor
NP: nurse practitioner
PA: physician assistant
TC: triage call

Edited by C Lovis; submitted 17.10.22; peer-reviewed by B McMillan, P Han; comments to author 22.11.22; revised version received 15.12.22; accepted 08.01.23; published 06.02.23.

Please cite as:
Ferguson K, Fraser M, Tuna M, Bruntz C, Dahrouge S
The Impact of an Electronic Portal on Patient Encounters in Primary Care: Interrupted Time-Series Analysis
JMIR Med Inform 2023;11:e43567
URL: https://medinform.jmir.org/2023/1/e43567
doi:10.2196/43567
PMID:36745495

Original Paper

An Electronic Dashboard to Improve Dosing of Hydroxychloroquine Within the Veterans Health Care System: Time Series Analysis

Anna Montgomery[1], MPH; Gary Tarasovsky[1,2], BSc; Zara Izadi[2], PharmD, PhD; Stephen Shiboski[2], PhD; Mary A Whooley[1,2,3], MSc, MD; Jo Dana[1], NP; Iziegbe Ehiorobo[2], MD; Jennifer Barton[4], MD; Lori Bennett[5], PharmD; Lorinda Chung[6,7], MD; Kimberly Reiter[8,9], MD; Elizabeth Wahl[10], MD; Meera Subash[11], MD; Gabriela Schmajuk[1,2,3], MSc, MD

[1]San Francisco VA Medical Center, San Francisco, CA, United States
[2]University of California San Francisco, San Francisco, CA, United States
[3]UCSF Philip R Lee Institute for Health Policy Studies, San Francisco, CA, United States
[4]VA Portland Health Care System, Portland, OR, United States
[5]Ralph H Johnson VA Medical Center, Charleston, SC, United States
[6]Palo Alto VA Medical Center, Palo Alto, CA, United States
[7]Stanford University, Palo Alto, CA, United States
[8]Raymond G Murphy VA Medical Center, Albuquerque, AZ, United States
[9]University of New Mexico School of Medicine, Albuquerque, AZ, United States
[10]Seattle/Puget Sound VA Healthcare System, Seattle, WA, United States
[11]UT Physicians Center for Autoimmunity, Houston, TX, United States

Corresponding Author:
Gabriela Schmajuk, MSc, MD
University of California San Francisco
4150 Clement St
San Francisco, CA, 94121
United States
Phone: 1 415 221 4810
Email: Gabriela.schmajuk@ucsf.edu

Abstract

Background: Hydroxychloroquine (HCQ) is commonly used for patients with autoimmune conditions. Long-term use of HCQ can cause retinal toxicity, but this risk can be reduced if high doses are avoided.

Objective: We developed and piloted an electronic health record–based dashboard to improve the safe prescribing of HCQ within the Veterans Health Administration (VHA). We observed pilot facilities over a 1-year period to determine whether they were able to improve the proportion of patients receiving inappropriate doses of HCQ.

Methods: Patients receiving HCQ were identified from the VHA corporate data warehouse. Using PowerBI (Microsoft Corp), we constructed a dashboard to display patient identifiers and the most recent HCQ dose and weight (flagged if ≥5.2 mg/kg/day). Six VHA pilot facilities were enlisted to test the dashboard and invited to participate in monthly webinars. We performed an interrupted time series analysis using synthetic controls to assess changes in the proportion of patients receiving HCQ ≥5.2 mg/kg/day between October 2020 and November 2021.

Results: At the start of the study period, we identified 18,525 total users of HCQ nationwide at 128 facilities in the VHA, including 1365 patients at the 6 pilot facilities. Nationwide, at baseline, 19.8% (3671/18,525) of patients were receiving high doses of HCQ. We observed significant improvements in the proportion of HCQ prescribed at doses ≥5.2 mg/kg/day among pilot facilities after the dashboard was deployed (–0.06; 95% CI –0.08 to –0.04). The difference in the postintervention linear trend for pilot versus synthetic controls was also significant (–0.06; 95% CI –0.08 to –0.05).

Conclusions: The use of an electronic health record–based dashboard reduced the proportion of patients receiving higher than recommended doses of HCQ and significantly improved performance at 6 VHA facilities. National roll-out of the dashboard will enable further improvements in the safe prescribing of HCQ.

(JMIR Med Inform 2023;11:e44455) doi:10.2196/44455

KEYWORDS

medical informatics; patient safety; health IT; hydroxychloroquine; dashboard; Veterans Health Administration; audit and feedback; electronic health record

Introduction

Hydroxychloroquine (HCQ) is among the most commonly used medications for patients with autoimmune conditions and received special attention in 2020 as a potential treatment for COVID-19, resulting in drug shortages for chronic users [1]. These drug shortages, combined with recent guidelines emphasizing toxicities associated with long-term use, highlighted the issue of prescribing HCQ in appropriate doses. Long-term use of HCQ, especially at higher doses, can cause severe retinal toxicity in some patients. The risk of this toxicity is reduced if the average daily dose of HCQ is ≤5 mg/kg/day [2,3]. However, recent studies have revealed that 30%-40% of patients prescribed HCQ receive doses >5 mg/kg/day [4,5].

Previous studies have shown that enterprise-wide national dashboards are capable of improving care, but they have not been developed quickly enough, or disseminated widely enough, to make meaningful, population-level impacts on process or outcome measures [6-10]. Local, electronic health record (EHR)–based medication safety dashboards have been used to support medication safety but have not been scaled to date [11-14].

In this study, we sought to develop and deploy a national EHR-based medication safety dashboard within the Veterans Health Administration (VHA) to reduce inappropriate HCQ dosing. The VHA is the largest integrated health care delivery system in the United States, serving over 9 million veterans nationwide. Six VHA pilot facilities were enlisted to test the dashboard and invited to participate in monthly webinars. We followed pilot facilities over a 1-year period to determine whether they were able to improve the proportion of patients receiving inappropriate doses of HCQ.

Methods

Dashboard Development

The dashboard was developed as part of an ongoing project to improve the safe prescribing of high-risk disease-modifying antirheumatic drugs among VHA patients by the San Francisco VA's Measurement Science Quality Enhancement Research Initiative. It was created using PowerBI (Microsoft Corp), a data management software package available within the VHA for approved users with secure access to EHR data. The VHA's corporate data warehouse (CDW), which contains national VHA EHR data, served as the data source for the dashboard (Multimedia Appendix 1). PowerBI allows developers to extract, analyze, and display data from a variety of sources and features interactive tables and graphs that can be filtered or expanded using a graphical user interface [15]. Notably, PowerBI dashboards are "read-only," that is, users can see and filter data elements, but in order to change data (eg, update the dose of HCQ), they must do so within the EHR.

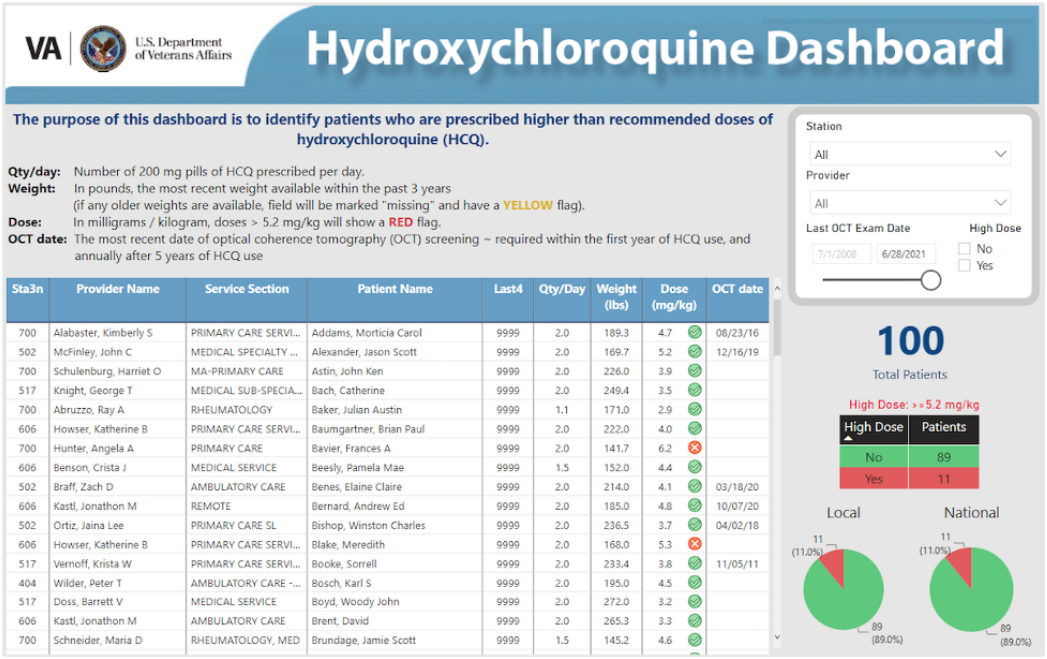
Facilities, Patients, and Data Elements

All 130 VHA facilities were eligible to be included in the study. We excluded 1 facility that had transitioned to the Cerner EHR and did not have patient data available in VistA, and 1 facility with fewer than 10 patients on HCQ, leaving 128 facilities for the analysis. Patients from included facilities in the VHA with a current, active prescription for HCQ were included in the data captured by the dashboard. Patients were excluded if the patient was deceased, or if the HCQ prescription indicated that it was a placebo or study drug. We extracted values for each patient's most recently prescribed HCQ dose (in mg per day), derived from the "quantity dispensed" and "days-supply" fields in the medication order. We also extracted the most recently captured body weight (in kg) to calculate the HCQ dose in mg/kg/day. These data were then linked to Microsoft PowerBI Gateway servers, which are automatically updated every 24 hours to reflect new information from CDW (Multimedia Appendix 1).

Dashboard Features

Figure 1 illustrates the dashboard using fictitious patient data. The dashboard displayed patient identifiers (first and last name, last 4 digits of their social security number, and VHA facility), the number of HCQ pills prescribed per day, most recently documented weight, date of most recent documented optical coherence tomography (OCT) exam, and prescriber name and service. Calculated fields included HCQ dose in mg/kg/day based on weight and the number of pills prescribed per day. Rows were marked with a red x-mark if the HCQ daily dose was calculated to be ≥5.2 mg/kg/day (vs a green check mark if <5.2 mg/kg/day) in the column immediately to the right of the dose. Patients without a recorded weight within the past 3 years were flagged with a yellow circle, indicating missing data. Rows could be filtered by facility location, provider, OCT exam date, or by HCQ dose. The dashboard also displayed national-, facility-, and prescriber-level performance (proportion of patients with HCQ doses ≥5.2 mg/kg/day out of the total number of patients receiving HCQ) shown as pie charts for benchmarking. A user guide and video tutorial for the dashboard were available via a web-based link on the dashboard landing page. User interactions (number of times the dashboard is accessed per authorized user) were tracked using the PowerBI Activity Log feature [16].

Figure 1. The hydroxychloroquine patient safety dashboard (using fictitious data). The dashboard was created using the Microsoft PowerBI software. Sta3n: unique medical facility codes for each VA station; Last 4: last 4 digits of a patient's social security number needed to identify a patient in the computerized patient record system (CPRS).



Study Period

Initial queries using CDW and PowerBI began in June 2020. The beginning of the study period—when baseline data collection on HCQ dosing across all VHA facilities started—began on August 11, 2020, prior to sharing the dashboard with any pilot testing facilities (see Multimedia Appendix 2). Of note, the final dashboard was developed over a period of under 5 months (June 2020 to October 2020).

Pilot Testing Facilities

We enlisted rheumatology providers, pharmacists, and dermatologists from 6 VHA pilot facilities to test the dashboard between October 26, 2020, and December 6, 2021. Pilot facilities were selected based on their willingness to participate in a related study involving screening for infections prior to immunosuppression. The 6 VHA pilot facilities included Ralph H Johnson VA Medical Center, Charleston, SC; Palo Alto VA Health Care System, Palo Alto, CA; VA Portland Health Care System, Portland, OR; Raymond G Murphy VA Medical Center, Albuquerque, NM; San Francisco VA Medical Center, San Francisco, CA; and Puget Sound/Seattle VA Healthcare System, Puget Sound, WA.

Pilot facilities were invited to use the dashboard via email. Once they agreed, they were granted secure access along with any additional staff at that facility. All facilities were trained in the use of the dashboard via a web-based webinar. Site personnel were invited to participate in web-based meetings of a Rheumatology Quality and Safety Workgroup to share feedback, address any barriers, and update information on the use of the dashboard, every other month. Each facility leader was also sent a quarterly facility-specific report via email with run charts depicting the proportion of patients on HCQ at doses ≥5.2 mg/kg/day and the number of times their facility accessed the dashboard during that quarter.

Each pilot facility was encouraged to develop an individualized workflow for use of the dashboard. For example, some facilities would check the dashboard weekly or monthly, while others used the downloadable report feature to distribute flagged patients to individual providers or trainees. All facility workflows included review of the dashboard, review of EHR charts of flagged patients, and HCQ dose adjustment if appropriate.

Control Facilities

Facilities in the control group did not have access to the dashboard and were not contacted as part of this study. Data on patients receiving HCQ were collected from CDW using the same process as was used for pilot facilities.

Complex Medication Instructions and Policy Change

On the dashboard, HCQ dose in mg/kg/day was calculated based on the number of pills prescribed and the patient's most recent weight. However, occasionally, HCQ orders had complex instructions (eg, "Take 2 pills Monday through Friday, and 1 pill on Saturdays and Sundays"), which resulted in miscalculations of the daily dose based on these fields. Two authors (AM and GS) reviewed 939 randomly selected charts and found 3% (28/939) of HCQ orders contained complex instructions. In order to reduce the chances of misclassifying patients as having an inappropriate HCQ dose due to complex instructions or fluctuating patient weights, on November 30, 2020, we made a policy change to designate doses of ≥5.2 mg/kg/day as "high dose" (as opposed to doses ≥5.0 mg/kg/day).

Covariates and Descriptive Variables

We assessed facility characteristic variables that might be important in relation to medication safety practices in general and HCQ dosing specifically: facility region (Midwest, North Atlantic, Continental, Southeast, and Pacific); facility

complexity (high, medium, and low); and the total number of patients prescribed HCQ at the facility [17].

In addition, we reported facility-level HCQ patient characteristics including the proportion of patients who were ≥55 years; self-identified non-Hispanic White, self-identified Hispanic or Latinx; with at least 1 VHA rheumatology clinic visit within 12 months of the beginning of the study period; and with a rural residence. Facility-level HCQ patient clinical factors included the proportion of patients with rheumatic diseases (rheumatoid arthritis, systemic lupus erythematosus [SLE], or other); with OCT exam documented; and the proportion with inappropriate HCQ dosing at baseline (August 11, 2020). A patient was considered to have a diagnosis if they had at least 2 codes (at least 30 days apart) for a specific condition listed here: rheumatoid arthritis, SLE, polymyalgia rheumatica, discoid lupus, nongout crystal arthropathy, undifferentiated connective tissue disease, sarcoidosis, antiphospholipid antibody syndrome, mixed connective tissue disease, systemic sclerosis, osteoarthritis, inflammatory myopathies (including polymyositis and dermatomyositis), psoriatic arthritis, ankylosing spondylitis, antineutrophil cytoplasmic antibody-associated vasculitis, other vasculitis (including Kawasaki disease), dermatitis, or giant cell arteritis.

Statistical Analysis

Descriptive statistics were used to summarize facility characteristics and facility-level patient characteristics. We used interrupted time series (ITS) analysis to assess the effects of the dashboard on observed changes in the proportion of patients with HCQ doses ≥5.2 mg/kg/day. ITS is a strong quasiexperimental study design that can be used for single- and multiple group comparisons. In an ITS analysis, the outcome variable of interest (eg, the average proportion of patients with HCQ doses ≥5.2 mg/kg/day) is observed over multiple time periods before and after an intervention that is expected to "interrupt" the trend over time. ITS has been previously found useful when evaluating health care interventions for its ability to evaluate the causal impact of policy changes and health care interventions without random assignment [18,19]. We used the *itsa* command, which is available in the official Stata packages *newey* and *prais* [19].

Due to large variability in key facility characteristics observed at baseline between pilot and control facilities (proportion of patients on HCQ at doses ≥5.2 mg/kg/day, facility complexity, and mean number of patients prescribed HCQ at the facility; ITS regression output are displayed in Multimedia Appendix 3), we opted to implement a robust matching method using synthetic controls to measure the impact of the dashboard on HCQ dosing at the pilot facilities. Using this approach, pilot facility performance was compared to matched synthetic controls using the *synth* package in Stata [20]. Synthetic controls were constructed from a weighted combination of control units not exposed to the dashboard but with preintervention outcome dynamics and covariate levels similar to the pilot facilities prior to any interventions [21]. Matching was based on observed changes in the proportion of patients with HCQ doses ≥5.2 mg/kg/day, facility complexity, and the mean number of patients prescribed HCQ at the facility. To assess the balance of the pilot facilities and their synthetic controls, we used the absolute standardized mean difference (ASMD). As a rule of thumb, ASMD < 0.10 is an indicator of a good balance between synthetic control unit and a treated unit [22].

As part of the multiple group ITS analysis, pilot facilities were compared to synthetic controls in weekly increments of the proportion of patients with HCQ doses ≥5.2 mg/kg/day. We estimated the coefficients using segmented ordinary least square (OLS) linear regression models in which the errors were assumed to follow a first-order autoregressive process [19]. The model was specified to base the pooled autocorrelation estimate on the autocorrelation of the residuals. We expressed the effect of the dashboard on the proportion of patients with HCQ doses ≥5.2 mg/kg/day as intercept and slope changes. The intervention date was set as October 26, 2020 (the date the pilot facilities were granted access to the dashboard). We incorporated the policy shift (shift from recording the proportion of patients receiving ≥5.0 mg/kg/day to those receiving ≥5.2 mg/kg/day on November 30, 2020) using established methods [18].

Statistical analyses were performed using Stata 15 (StataCorp LLC). A P value <.05 was used as the criterion for statistical significance.

Secondary Analyses

As secondary analyses, we compared the 6 pilot facilities to other facilities using modified Xbar-R charts. We used Microsoft QI Macros, a statistical process control software package plugin for Microsoft Excel, to generate modified Xbar-R charts to analyze the overall trends and stability in the proportion of patients with HCQ doses ≥5.2 mg/kg/day over time. Upper and lower control limits varied based on the average proportion of patients with HCQ doses ≥5.2 mg/kg/day. A continuous change of 6 or more points in a row or 8 or more points on the same side of the centerline is considered a significant trend [23].

We performed 2 separate comparisons: (1) pilot facilities versus all other facilities nationally and (2) pilot facilities versus matched control facilities. Matched control facilities were selected based on (1) the slope of proportion of patients prescribed HCQ at doses ≥5.2 mg/kg/day during the baseline period (August 11, 2020, to October 25, 2020); (2) the total number of patients prescribed HCQ; and (3) high facility complexity. Since pilot facilities had a mean) of 228 (SD 69) patients prescribed HCQ, we required matched control facilities to have at least 75 patients prescribed HCQ. Application of these criteria resulted in 8 matched control facilities, which were all included in the matched control sensitivity analysis.

Feedback From Pilot Facilities

At the end of the study period, clinicians at pilot facilities were sent a confidential survey to solicit quantitative and qualitative feedback about the dashboard. The 14-item survey included questions about the capacity in which sites used the dashboard, usability of the dashboard, suggestions for improvement, and the likelihood of recommending the dashboard to a colleague or trainee.

Ethics Approval

All VHA authors of this manuscript attest that the activities that resulted in producing this manuscript were conducted as part of a nonresearch evaluation under the authority of the National Rheumatology Field Advisory Committee and Center for Medication Safety. This work was approved by the VA Quality Enhancement Research Initiative (QUERI; IRB 15-18358).

Results

Pilot Facilities and Workflows

We identified 18,525 total users of HCQ nationwide in the VHA, including 1365 patients at the 6 pilot facilities. Across the 6 pilot facilities, 36 providers were granted access to the dashboard including 14 rheumatologists, 12 physician residents, 3 rheumatology fellows, 2 nurse practitioners specializing in rheumatology, 2 clinical pharmacists, 1 dermatologist, 1 registered nurse coordinator, and 1 primary care physician. Different pilot facilities developed different workflows around dashboard use to suit their needs. Some facilities requested access for all their clinicians (attendings and trainees) and had each one review their own patients. Others had a designated reviewer who checked the dashboard once a month or once a quarter. Another was able to download a spreadsheet containing the dashboard data and distribute it securely for clinician review (an example of typical dashboard clinic workflow for users is available in Multimedia Appendix 4). All pilot facilities had at least 20 interactions with the dashboard starting in October 2020; the median weekly number of dashboard interactions over the course of the study period was 8 (IQR 4-15).

Baseline Facility Characteristics

Table 1 shows the characteristics of pilot facilities at all facilities nationally. Nationwide, at the start of the study period, 19.8% (3671/18,525; range 4.26% to 44%) of patients prescribed HCQ were receiving HCQ ≥5 mg/kg/day versus 16.1% (220/1365) among pilot facilities.

Table 1. Facility characteristics and practice-level patient characteristics for the pilot versus all facilities at baseline (November 8, 2020).

Facility characteristics	Pilot facilities (n=6)	All facilities (N=128)
Complexity[a], n (%)		
High complexity	6 (100)	84 (66)
Medium complexity	0 (0)	18 (14)
Low complexity	0 (0)	26 (20)
Geographic location, n (%)		
Continental	0 (0)	24 (19)
Midwest	0 (0)	26 (20)
North Atlantic	0 (0)	36 (28)
Pacific	5 (83)	22 (17)
Southeast	1 (17)	20 (16)
Total patients prescribed HCQ[b], mean (SD)	228 (69)	146 (107)
Facility-level HCQ patient characteristics, mean (SD)		
Proportion of male patients	0.72 (0.05)	0.71 (0.10)
Proportion of patients aged >55 years	0.79 (0.10)	0.76 (0.09)
Proportion of non-White patients	0.33 (0.10)	0.32 (0.18)
Proportion of Hispanic/Latinx patients	0.05 (0.03)	0.05 (0.04)
Proportion of patients who visited a VA rheumatology clinic within 1 year of baseline	0.44 (0.08)	0.59 (0.19)
Proportion of patients with a rural residence	0.33 (0.02)	0.32 (0.21)
Facility-level HCQ patient clinical factors, mean (SD)		
Proportion of patients with rheumatoid arthritis[c]	0.43 (0.15)	0.45 (0.10)
Proportion of patients with systemic lupus erythematosus[c]	0.14 (0.02)	0.16 (0.04)
Proportion of patients with other rheumatic disease[c]	0.20 (0.12)	0.20 (0.12)
Proportion of patients with HCQ dose ≥5 mg/kg/day at baseline (August 11, 2020)	0.16 (0.03)	0.20 (0.07)

[a]Station complexity: high complexity facilities have large levels of patient volume, patient risk, teaching and research, and contain level 4 to 5 intensive care units; medium complexity facilities have medium levels of patient volume, medium patient risk, some teaching and/or research, and contain level 3 and 4 intensive care units; low complexity facilities have the smallest level of patient volume, little or no teaching/research, the lowest number of physician specialists per patient, and contain level 1 and 2 intensive care units.

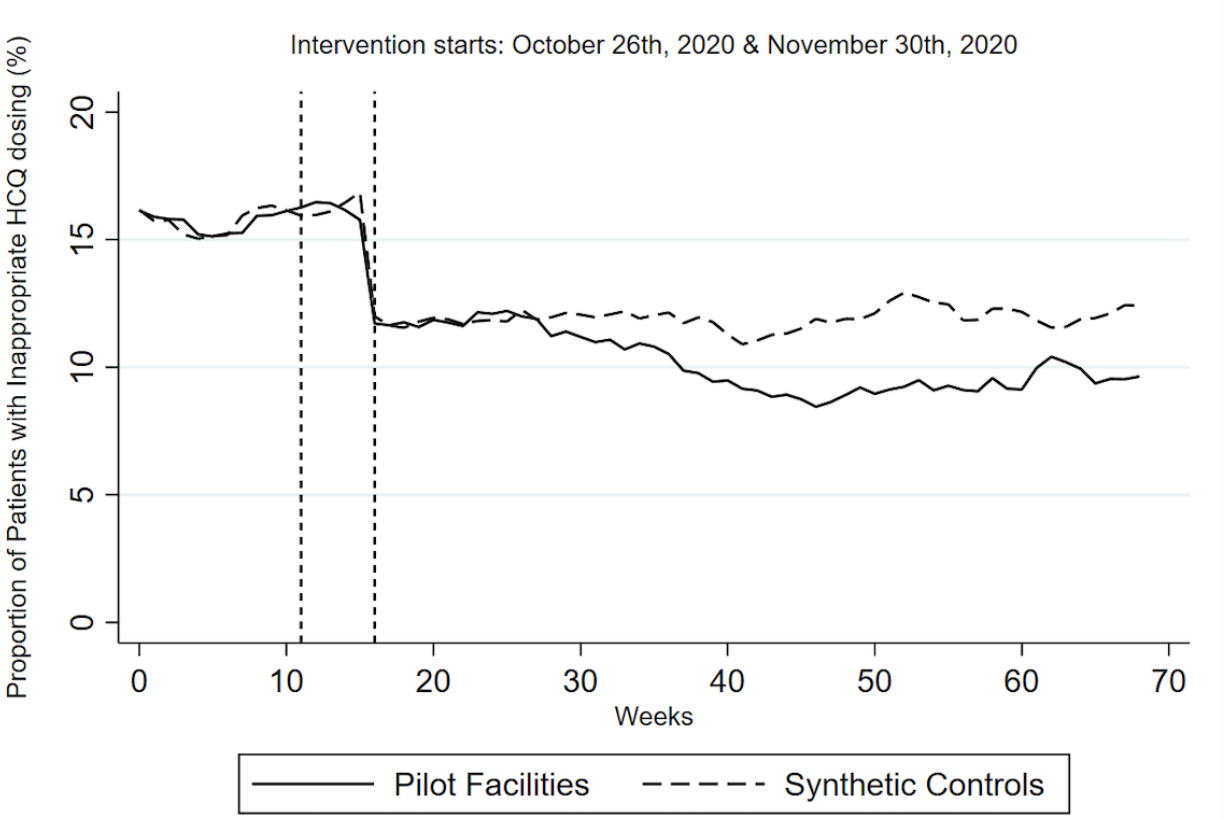
[b]HCQ: hydroxychloroquine.

[c]Rheumatic diseases were identified as veterans with 2 or more ICD-10 codes within the same disease category, separated by 30 or more days. Other autoimmune rheumatic diseases included: polymyalgia rheumatica, discoid lupus erythematosus, nongout crystal arthropathy, undifferentiated connective tissue disease, sarcoidosis, antiphospholipid syndrome, mixed connective tissue disease, systemic sclerosis, osteoarthritis, inflammatory myopathies (including polymyositis and dermatomyositis), psoriatic arthritis, ankylosing spondylitis, antineutrophil cytoplasmic antibody-associated vasculitis, and other vasculitis (including Kawasaki disease), lymphocytic infiltrates of the skin, or giant cell arteritis.

ITS Analysis With Synthetic Controls

Pilot facilities and synthetic controls were well matched in their predictor balance (ASMD=0.05). The postintervention linear trend showed pilot facilities' proportion of patients with HCQ doses ≥5.2 mg/kg/day changed by –0.06 (95% CI –0.08 to –0.04) after the policy change, while the synthetic controls remained stable (0.006; 95% CI –0.00 to 0.01), with a statistically significant difference between the 2 groups by the end of the study period (–0.06; 95% CI –0.08 to –0.05; Multimedia Appendix 5 and Figure 2).

Figure 2. Synthetic control analysis with mean proportion of patients with high HCQ doses among the 6 pilot facilities and synthetic controls. The first intervention (October 26, 2020) was the date on which the dashboard was shared with the 6 pilot facilities. The second intervention (November 30, 2020) captured the policy change of adjusting the "high dose" definition ≥5.0 mg/kg/day to ≥5.2mg/kg/day to account for complex prescription instructions. HCQ: hydroxychloroquine.



Secondary Analyses

As seen in Multimedia Appendix 6, the modified Xbar-R control chart showed meaningful improvements in the proportion of patients receiving HCQ doses ≥5.2 mg/kg among pilot facilities over the course of the study period. There was a downward trend of 21 points outside of the upper and lower control limits, indicating a significant overall average process change. In contrast, the 8 matched control facilities' proportion remained stable (ie, within the control limits). A comparison of pilot facilities to all other facilities nationally revealed similar results (Multimedia Appendix 7).

Feedback From Pilot Sites

Six clinicians, 1 from each pilot facility, responded to the web-based survey. Of these 6 clinicians, 5 reported that the dashboard was extremely easy to use, 5 answered they were extremely likely to use the dashboard in the future, and 5 responded they were extremely likely to recommend the dashboard to a colleague or trainee.

Discussion

In an era where the complexity of care and the number of evidence-based practices are ever expanding, the cognitive load required to address these practices during a short office visit can be overwhelming for clinicians. EHR-based dashboards are 1 method to support clinicians in evidence-based care of their patients. In this study, we developed an EHR-based medication safety dashboard to improve the safe prescribing of HCQ within the VHA. As part of a multipronged intervention, we found that audit and feedback via the dashboard resulted in a clinically meaningful and statistically significant reduction in the proportion of patients receiving high doses of HCQ among pilot facilities. Based on our linear postintervention trends, on the assumption that all facilities will behave similarly to the pilot sites, it would take approximately 4 years to reduce the proportion of patients receiving high HCQ doses from 16% to less than 5%.

Several features of the infrastructure available through the VHA made this a successful pilot. First, enterprise-wide PowerBI software was easily accessible as a pre-existing software suite available within the VHA for internal users. A new workspace was requested and granted within 48 hours; no new software installation was required. Second, it was straightforward to query VHA CDW data and then link these data to PowerBI servers. Construction of a first prototype of the dashboard took only a few months, and the final version (after several iterations) was available in 5 months. Beyond the VHA infrastructure, this pilot was feasible because of its limited scope to a single medication—HCQ comes in a single pill size, and most prescriptions have the number of pills dispensed corresponding to the daily dose, which facilitated calculating dose in mg/kg/day. Finally, because of its intuitive user interface, training required to use the dashboard by pilot sites was minimal.

There are few descriptions of EHR-integrated medication safety dashboards in the literature, and those that have been reported

have also been successful [24]. For example, with the use of the UK SMASH dashboard, the prevalence of potentially unsafe prescribing of nonsteroidal anti-inflammatory drugs and other medications was reduced by 41% at intervention facilities [10,11]. Another, US-based, local, pharmacist-led medication safety program, which included a dashboard and educational outreach, reduced errors by 27%-49% after 6 and 12 months of use [25]. Several US patient registries have also developed clinician-facing dashboards to improve quality and medication safety and demonstrated significant improvements over time [26].

Although our pilot project was successful, we do note some limitations for the dashboard. First, although the development and validation of data in the dashboard were smooth, there were a small fraction of patients whose calculated doses remained inaccurate due to complex instructions that did not match the number of pills supplied. We attempted to mitigate these inaccuracies by only flagging doses ≥5.2 mg/kg/day instead of 5 mg/kg/day. We made this choice to avoid falsely labeling HCQ orders as high at the expense of missing some patients receiving doses above those recommended in the latest guidelines. Further work is needed to explore whether this tradeoff is worthwhile, especially since many clinicians use complex dosing in order to avoid average daily doses of ≥5 mg/kg/day, so the use of complex instructions may be correlated with appropriate dosing. In the future, 1 potential solution could be to develop an algorithm that captures information from the complex instructions using natural language processing techniques. Second, many pilot users requested additional features that are not available through PowerBI. Most importantly, users wished to be able to annotate dashboard tables directly or unflag patients who might be receiving higher than recommended doses of HCQ deliberately due to severe disease. Unfortunately, these features were not available in the VHA implementation of PowerBI at the time of this study.

One critical question for the future is whether the improvements observed in this pilot study will be sustainable. Clinician buy-in and ongoing utilization are crucial to the effectiveness of this dashboard as a sustainable audit-and-feedback tool [25]. Several of our pilot facilities started using this dashboard as a component of their routine quality improvement activities and reported dashboard use as part of a pay-for-performance program. Other facilities incorporated its use into trainee quality improvement activities. These additional use cases make sustainability more likely.

Another important question for future studies is about the clinical effects of reducing HCQ doses for some patients. Some recent observational studies have suggested that patients with SLE who decrease their HCQ dose may be at increased risk for disease flares [27,28]. It seems unlikely that small changes in dosing would have a large effect, but nevertheless, this is an important question to investigate. Unfortunately, since this is a national study limited by using structured EHR data, it is impossible to ascertain the condition of any specific patient before or after introducing the dashboard.

Moving forward, we will test the effects of the dashboard in a national roll-out across all VHA facilities. Additional mixed methods research will aid our understanding of provider adoption and sustained use of the dashboard and whether other interventions are needed to support safe prescribing of HCQ (eg, clinical decision support for weight-based dosing, or other pharmacy-based alerts or workflows) [29]. We also plan to roll-out additional dashboards focused on other important rheumatology safety issues, including pretreatment screening for latent infections in patients receiving biologic and targeted small molecule medications and HLA B:5801 testing for eligible patients receiving allopurinol. Our hope is that with a suite of dashboards and associated toolkits, quality improvement activities will be more feasible for all clinicians.

In summary, we successfully developed and deployed an EHR-based medication safety dashboard to improve the safe prescribing of HCQ within the VHA. The use of the dashboard significantly reduced the proportion of patients receiving higher than recommended doses of HCQ at 6 VHA facilities. National roll-out of the dashboard will enable further improvements in the safe prescribing of HCQ.

Acknowledgments

This work was funded by the VA Quality Enhancement Research Initiative (QUERI) grant QIS 19-459.

Data Availability

The data that support the findings of this study are available from the Department of Veterans Affairs, but restrictions apply to the availability of these data. Data are, however, available from the authors upon reasonable request and with permission of the Department of Veterans Affairs.

Authors' Contributions

AM made substantial contributions to the conception, design, data analysis and interpretation, and the creation of the dashboards that were used in this work. GT contributed to the study design, the creation of the dashboard, and assisted with the interpretation of the data used in this work. SS and ZI made substantial contributions to the data analysis and interpretation of the data along with assistance with the drafting of the manuscript. MAW, JD, IE, JB, LB, LC, KR, EW, and MS made substantial contributions to the data acquisition and the revisions to the manuscript. GS made substantial contributions to the conception, study design, data interpretation, and drafting of the manuscript for submission. All authors read and approved the final manuscript.

Conflicts of Interest

ZI is an employee at BMS.

Multimedia Appendix 1
Architecture diagram of the hydroxychloroquine patient safety dashboard. The electronic health record data were pulled from the corporate data warehouse SQL Servers. The data were then directly linked from the SQL Servers to the PowerBI gateway and presented to the end user via a web interface.
[DOCX File , 77 KB - medinform_v11i1e44455_app1.docx]

Multimedia Appendix 2
Study timeline.
[DOCX File , 13 KB - medinform_v11i1e44455_app2.docx]

Multimedia Appendix 3 [DOCX File , 15 KB - medinform_v11i1e44455_app3.docx]

Multimedia Appendix 4
Hydroxychloroquine patient safety dashboard pilot facility Swimlane diagram.
[DOCX File , 176 KB - medinform_v11i1e44455_app4.docx]

Multimedia Appendix 5
Comparison of a linear postintervention trend for pilot versus synthetic control pilot facilities after the policy change (November 30, 2020).
[DOCX File , 14 KB - medinform_v11i1e44455_app5.docx]

Multimedia Appendix 6
Modified Xbar control chart showing mean percent of patients receiving inappropriate HCQ doses among pilot facilities and matched control facilities. The x-axis shows 2-week time segments during the study period from August 11, 2020, to December 6, 2021; the y-axis shows the percent of patients with higher than recommended HCQ doses. The vertical dotted lines denote the dates when the pilot facilities were granted access to the dashboard and when the “High Dose HCQ” definition changed from ≥5.0 mg/kg/day to ≥5.2 mg/kg/day. The orange and blue dotted lines show the average and upper/lower control limits for the 6 pilot facilities compared to 8 matched control facilities, respectively. The dots on the solid lines represent performance at each time point. CL: central line; LCL: lower control limit; HCQ: hydroxychloroquine; UCL: upper control limit.
[DOCX File , 37 KB - medinform_v11i1e44455_app6.docx]

Multimedia Appendix 7
Modified Xbar control chart showing mean percent of patients receiving inappropriate HCQ doses among pilot facilities and all other facilities nationally. The x-axis shows 2-week time segments during the study period from August 11, 2020, to December 6, 2021; the y-axis shows the percent of patients with higher than recommended HCQ doses. The vertical dotted lines denote the dates when the pilot facilities were granted access to the dashboard and when the “High Dose HCQ” definition changed from ≥5.0 mg/kg/day to ≥5.2 mg/kg/day. The orange and blue dotted lines show the average and upper/lower control limits for the 6 pilot facilities compared to all other 122 facilities, respectively. The dots on the solid lines represent performance at each time point. CL: central line; LCL: lower control limit; HCQ: hydroxychloroquine; UCL: upper control limit.
[DOCX File , 37 KB - medinform_v11i1e44455_app7.docx]

References

1. Sattui SE, Liew JW, Graef ER, Coler-Reilly A, Berenbaum F, Duarte-García A, et al. Swinging the pendulum: lessons learned from public discourse concerning hydroxychloroquine and COVID-19. Expert Rev Clin Immunol 2020;16(7):659-666 [FREE Full text] [doi: 10.1080/1744666X.2020.1792778] [Medline: 32620062]
2. Melles RB, Marmor MF. The risk of toxic retinopathy in patients on long-term hydroxychloroquine therapy. JAMA Ophthalmol 2014;132(12):1453-1460. [doi: 10.1001/jamaophthalmol.2014.3459] [Medline: 25275721]
3. Rosenbaum JT, Costenbader KH, Desmarais J, Ginzler EM, Fett N, Goodman SM, et al. American College of Rheumatology, American Academy of Dermatology, Rheumatologic Dermatology Society, and American Academy of Ophthalmology 2020 joint statement on hydroxychloroquine use with respect to retinal toxicity. Arthritis Rheumatol 2021;73(6):908-911. [doi: 10.1002/art.41683] [Medline: 33559327]
4. Gianfrancesco MA, Schmajuk G, Haserodt S, Trupin L, Izadi Z, Jafri K, et al. Hydroxychloroquine dosing in immune-mediated diseases: implications for patient safety. Rheumatol Int 2017;37(10):1611-1618 [FREE Full text] [doi: 10.1007/s00296-017-3782-6] [Medline: 28748425]

5. Izadi Z, Gianfrancesco M, Evans M, Kay J, Trupin L, Schmajuk G, et al. OPO263 hydroxychloroquine dosing in patients with rheumatic disease across the U.S.: data from the rheumatology informatics system for effectiveness (RISE) registry. Ann Rheum Dis 2019;78(2):212-213.
6. Khairat SS, Dukkipati A, Lauria HA, Bice T, Travers D, Carson SS. The impact of visualization dashboards on quality of care and clinician satisfaction: integrative literature review. JMIR Hum Factors 2018;5(2):e22 [FREE Full text] [doi: 10.2196/humanfactors.9328] [Medline: 29853440]
7. Burningham Z, Jackson GL, Kelleher J, Stevens M, Morris I, Cohen J, et al. The enhancing quality of prescribing practices for older veterans discharged from the emergency department (EQUIPPED) potentially inappropriate medication dashboard: a suitable alternative to the in-person academic detailing and standardized feedback reports of traditional EQUIPPED? Clin Ther 2020;42(4):573-582. [doi: 10.1016/j.clinthera.2020.02.013] [Medline: 32222360]
8. Foster M, Albanese C, Chen Q, Sethares KA, Evans S, Lehmann LS, et al. Heart failure dashboard design and validation to improve care of veterans. Appl Clin Inform 2020;11(1):153-159. [doi: 10.1055/s-0040-1701257] [Medline: 32102107]
9. Lau MK, Bounthavong M, Kay CL, Harvey MA, Christopher MLD. Clinical dashboard development and use for academic detailing in the U.S. Department of Veterans Affairs. J Am Pharm Assoc (2003) 2019;59(2S):S96-S103.e3. [doi: 10.1016/j.japh.2018.12.006] [Medline: 30713078]
10. Fischer MJ, Kourany WM, Sovern K, Forrester K, Griffin C, Lightner N, et al. Development, implementation and user experience of the veterans health administration (VHA) dialysis dashboard. BMC Nephrol 2020;21(1):136 [FREE Full text] [doi: 10.1186/s12882-020-01798-6] [Medline: 32299383]
11. Linder JA, Schnipper JL, Tsurikova R, Yu DT, Volk LA, Melnikas AJ, et al. Electronic health record feedback to improve antibiotic prescribing for acute respiratory infections. Am J Manag Care 2010;16(suppl 12 HIT):e311-e319. [Medline: 21322301]
12. Koopman RJ, Kochendorfer KM, Moore JL, Mehr DR, Wakefield DS, Yadamsuren B, et al. A diabetes dashboard and physician efficiency and accuracy in accessing data needed for high-quality diabetes care. Ann Fam Med 2011;9(5):398-405 [FREE Full text] [doi: 10.1370/afm.1286] [Medline: 21911758]
13. Fletcher GS, Aaronson BA, White AA, Julka R. Effect of a real-time electronic dashboard on a rapid response system. J Med Syst 2017;42(1):5. [doi: 10.1007/s10916-017-0858-5] [Medline: 29159719]
14. Taber DJ, Pilch NA, McGillicuddy JW, Mardis C, Treiber F, Fleming JN. Using informatics and mobile health to improve medication safety monitoring in kidney transplant recipients. Am J Health Syst Pharm 2019;76(15):1143-1149. [doi: 10.1093/ajhp/zxz115] [Medline: 31361870]
15. Data sources in Power BI desktop. Microsoft. URL: https://docs.microsoft.com/en-us/power-bi/connect-data/desktop-data-sources [accessed 2022-01-20]
16. Track user activities in Power BI. Microsoft. URL: https://docs.microsoft.com/en-us/power-bi/admin/service-admin-auditing [accessed 2022-01-20]
17. Site facility name and complexity. United States Department of Veterans Affairs. URL: https://tinyurl.com/4b4s68zy [accessed 2022-01-20]
18. Riley WT, Glasgow RE, Etheredge L, Abernethy AP. Rapid, responsive, relevant (R3) research: a call for a rapid learning health research enterprise. Clin Transl Med 2013;2(1):10 [FREE Full text] [doi: 10.1186/2001-1326-2-10] [Medline: 23663660]
19. Linden A. Conducting interrupted time-series analysis for single- and multiple-group comparisons. Stata J 2015;15(2):480-500 [FREE Full text]
20. Abadie A, Diamond A, Hainmueller J. Synthetic control methods for comparative case studies: estimating the effect of California's tobacco control program. J Am Stat Assoc 2010;105(490):493-505. [doi: 10.1198/jasa.2009.ap08746]
21. Kreif N, Grieve R, Hangartner D, Turner AJ, Nikolova S, Sutton M. Examination of the synthetic control method for evaluating health policies with multiple treated units. Health Econ 2016;25(12):1514-1528 [FREE Full text] [doi: 10.1002/hec.3258] [Medline: 26443693]
22. Parast L, Hunt P, Griffin BA, Powell D. When is a match sufficient? A score-based balance metric for the synthetic control method. J Causal Inference 2020;8(1):209-228 [FREE Full text]
23. Carey RG, Staker LV. Improving Healthcare with Control Charts: Basic and Advanced SPC Methods and Case Studies. Milwaukee, Wisconsin: ASQ Quality Press; 2002.
24. Dowding D, Randell R, Gardner P, Fitzpatrick G, Dykes P, Favela J, et al. Dashboards for improving patient care: review of the literature. Int J Med Inform 2015;84(2):87-100 [FREE Full text] [doi: 10.1016/j.ijmedinf.2014.10.001] [Medline: 25453274]
25. Avery AJ, Rodgers S, Cantrill JA, Armstrong S, Cresswell K, Eden M, et al. A pharmacist-led information technology intervention for medication errors (PINCER): a multicentre, cluster randomised, controlled trial and cost-effectiveness analysis. Lancet 2012;379(9823):1310-1319 [FREE Full text] [doi: 10.1016/S0140-6736(11)61817-5] [Medline: 22357106]
26. Izadi Z, Schmajuk G, Gianfrancesco M, Subash M, Evans M, Trupin L, et al. Significant gains in rheumatoid arthritis quality measures among RISE registry practices. Arthritis Care Res (Hoboken) 2022;74(2):219-228. [doi: 10.1002/acr.24444] [Medline: 32937026]

27. Almeida-Brasil CC, Hanly JG, Urowitz M, Clarke AE, Ruiz-Irastorza G, Gordon C, et al. Flares after hydroxychloroquine reduction or discontinuation: results from the systemic lupus international collaborating clinics (SLICC) inception cohort. Ann Rheum Dis 2022;81(3):370-378 [FREE Full text] [doi: 10.1136/annrheumdis-2021-221295] [Medline: 34911705]
28. Jorge AM, Mancini C, Zhou B, Ho G, Zhang Y, Costenbader K, et al. Hydroxychloroquine dose per ophthalmology guidelines and the risk of systemic lupus erythematosus flares. JAMA 2022;328(14):1458-1460. [doi: 10.1001/jama.2022.13591] [Medline: 36112387]
29. Gianfrancesco M, Murray S, Evans M, Schmajuk G, Yazdany J. A pragmatic randomized trial to improve safe dosing of hydroxychloroquine [abstract]. Arthritis Rheumatology. 2019. URL: https://acrabstracts.org/abstract/a-pragmatic-randomized-trial-to-improve-safe-dosing-of-hydroxychloroquine/ [accessed 2023-04-21]

Abbreviations

CDW: corporate data warehouse
EHR: electronic health record
HCQ: hydroxychloroquine
ITS: interrupted time series
OCT: optical coherence tomography
OLS: ordinary least square
SLE: systemic lupus erythematosus
VHA: Veterans Health Administration

Edited by C Lovis; submitted 21.11.22; peer-reviewed by H Wang, R Marshall; comments to author 27.01.23; revised version received 03.02.23; accepted 19.03.23; published 12.05.23.

Please cite as:
Montgomery A, Tarasovsky G, Izadi Z, Shiboski S, Whooley MA, Dana J, Ehiorobo I, Barton J, Bennett L, Chung L, Reiter K, Wahl E, Subash M, Schmajuk G
An Electronic Dashboard to Improve Dosing of Hydroxychloroquine Within the Veterans Health Care System: Time Series Analysis
JMIR Med Inform 2023;11:e44455
URL: https://medinform.jmir.org/2023/1/e44455
doi:10.2196/44455
PMID:37171858

Original Paper

Integrated Personal Health Record in Indonesia: Design Science Research Study

Nabila Clydea Harahap[1], Dr; Putu Wuri Handayani[1], Dr; Achmad Nizar Hidayanto[1], Prof Dr

Faculty of Computer Science, University of Indonesia, Depok, Indonesia

Corresponding Author:
Nabila Clydea Harahap, Dr
Faculty of Computer Science
University of Indonesia
Kampus UI Depok, Pondok Cina, Beji
Depok, 16424
Indonesia
Phone: 62 8571652699
Email: nabila.clydea@ui.ac.id

Abstract

Background: Personal health records (PHRs) are consumer-centric tools designed to facilitate the tracking, management, and sharing of personal health information. PHR research has mainly been conducted in high-income countries rather than in low- and middle-income countries. Moreover, previous studies that proposed PHR design in low- and middle-income countries did not describe integration with other systems, or there was no stakeholder involvement in exploring PHR requirements.

Objective: This study developed an integrated PHR architecture and prototype in Indonesia using design science research. We conducted the research in Indonesia, a low- to middle-income country with the largest population in Southeast Asia and a tiered health system.

Methods: This study followed the design science research guidelines. The requirements were identified through interviews with 37 respondents from health organizations and a questionnaire with 1012 patients. Afterward, the proposed architecture and prototype were evaluated via interviews with 6 IT or eHealth experts.

Results: The architecture design refers to The Open Group Architecture Framework version 9.2 and comprises 5 components: architecture vision, business architecture, application architecture, data architecture, and technology architecture. We developed a high-fidelity prototype for patients and physicians. In the evaluation, improvements were made to add the stakeholders and the required functionality to the PHR and add the necessary information to the functions that were developed in the prototype.

Conclusions: We used design science to illustrate PHR integration in Indonesia, which involves related stakeholders in requirement gathering and evaluation. We developed architecture and application prototypes based on health systems in Indonesia, which comprise routine health services, including disease treatment and health examinations, as well as promotive and preventive health efforts.

(JMIR Med Inform 2023;11:e44784) doi:10.2196/44784

KEYWORDS

personal health record; integrated; Indonesia; design science; mobile phone

Introduction

Background

Current trends in health informatics encourage the transition from institution-centric to patient-centric health care [1]. The use of IT is not only for patients in health care settings but also for all individuals who want to maintain health and are involved in disease prevention and health promotion [1]. Personal health records (PHRs) are consumer-centric tools designed to facilitate the tracking, management, and sharing of personal health information [1]. PHRs contain medical data and information about a patient that are managed by the patients themselves [2]. PHRs form a trend from information controlled by the health system to information controlled by individuals [3].

In its simplest form, a PHR is a stand-alone application (stand-alone PHR) and is not connected to other systems [4]. In a more complex form, the health information provided by the PHR is linked to the electronic health record or electronic

medical record (tethered PHR) [4]. Furthermore, PHRs can be connected to various health data sources to obtain and transmit data (integrated PHR) [4]. An integrated PHR is the most ideal form of PHR as implementing PHRs in this way has the potential to improve the quality, accessibility, and delivery of health services [3].

A previous review on the implementation of PHRs shows that PHR research has mainly been conducted in high-income countries rather than in low- and middle-income countries [5]. Few studies that have been conducted in low- and middle-income countries aim to propose PHR applications for certain purposes, such as pediatric vaccination [6], or specific diseases, such as metabolic syndrome management [7], chronic heart failure [8], and kidney transplant [9]. These studies focused on the usability of PHRs and did not describe integration with other systems or applications. A study by Abdulnabi et al [10] described PHR interoperability by designing a distributed PHR model. However, there was no stakeholder involvement in exploring PHR requirements.

Using design science, we complement gaps from previous studies by developing a PHR model that is integrated with various systems, and we involve relevant stakeholders to explore the requirements and evaluate the proposed PHR model. We conducted the research in Indonesia, a low- to middle-income country with the largest population in Southeast Asia [11,12]. In Indonesia, health services are delivered by the public and private sectors. In the public sector, health facilities comprise hospitals (general and specialty) and *pusat kesehatan masyarakat* or primary health centers (Puskesmas). In the private sector, health facilities comprise hospitals and primary care clinics. In addition, there is the Social Security Agency for Health or *Badan Pelaksana Jaminan Sosial Kesehatan* (BPJS Kesehatan) that administers the national health insurance program (*Jaminan Kesehatan Nasional* [JKN] or national health insurance). Patients with JKN must follow a tiered referral flow starting from primary care facilities as gatekeepers for JKN patients before being referred to hospitals. Without a referral letter, JKN patients are not allowed to go directly to a hospital or specialist clinic except in an emergency [13].

In addition to health efforts that focus on treating and curing diseases, there are also Healthy Family (Keluarga Sehat) and Community Healthy Life Movement (Germas) programs that are managed by the Puskesmas and focus on promotive and preventive health efforts [14,15]. Currently, health development policies in Indonesia are directed at improving access to and quality of health services, with an emphasis on increasing promotive and preventive health efforts supported by innovation and the use of technology [16]. Integrated PHRs can be an opportunity to improve access to and quality of health services in Indonesia by using IT [17].

Objectives

As a technological solution for integrated PHRs in Indonesia, this study developed an integrated PHR architecture and prototype in Indonesia using the design science research (DSR) approach by Hevner et al [18]. The DSR approach was chosen in this study as the goal of DSR is to focus on designing systems that not only are practical but can also contribute to knowledge. The question that will be answered in this research is as follows: How are the architectural designs and prototypes of integrated PHR applications in Indonesia? The design of the PHR model, which was developed using a DSR approach, can provide an overview for developing PHRs using scientific theory and methods. The results of this study are expected to be a guide for health facilities or health policy makers in integrating PHRs and health applications in Indonesia.

Methods

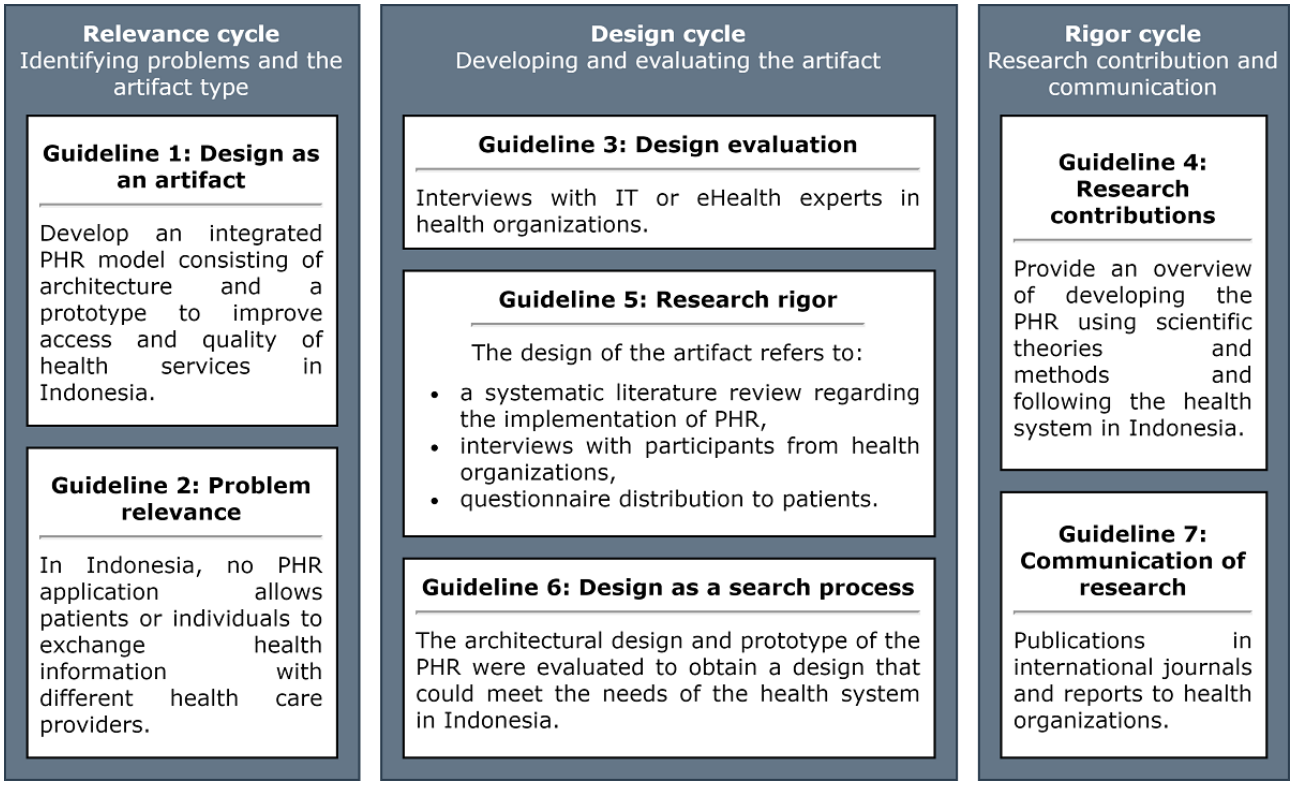
Design Science

This study was conducted using a DSR approach [18]. In DSR, the details or stages of design and development may vary even though the focus of the research is on artifact design [19]. Peffers et al [20] describe 6 activity-based methodologies for DSR, whereas Hevner et al [18] define 7 guidelines for DSR. This study follows the DSR guidelines defined by Hevner et al [18], which are grouped into 3 cycles or phases comprising the relevance cycle (identifying problems and the artifact type), design cycle (developing and evaluating the artifact), and rigor cycle (research contribution and communication; Figure 1 [5]). These guidelines have been followed by DSR studies in low- and middle-income countries to develop health information systems such as mobile health (mHealth) [19,21-27]. The guidelines by Hevner et al [18], as a necessary element in DSR, are also consistent with the DSR methodology by Peffers et al [20].

The artifacts developed in this study are architecture and application prototypes. Integrated PHR requirements were carried out based on a systematic literature review regarding functionalities and issues in the implementation of PHRs [5], literature studies on health regulations in Indonesia, interviews with health organizations, and questionnaire distribution to patients. The requirements mentioned by health organizations in the interviews were categorized and grouped into themes that were defined by Harahap et al [5] and combined with requirements from patients. The architectural design then became a reference for designing the application prototype.

This study used a purposive sampling method to select participants who had the required knowledge regarding PHR implementation. The health organizations from which participants were interviewed were health facilities (private and government hospitals, Puskesmas, and clinics), health regulators (Ministry of Health and BPJS Kesehatan), and health application vendors. Data were collected through semistructured interviews between August 19, 2020, and January 15, 2021, and between October 1, 2021, and October 11, 2021. Interviews were conducted on the web using Zoom Cloud Meetings (Zoom Video Communications) and audio recorded with the participants' consent. Each interview lasted between 30 and 60 minutes. The interview questions are attached in Multimedia Appendix 1.

Figure 1. Design science research guideline [5,18]. PHR: personal health record.



The questionnaire was distributed to patients who met the criteria of respondents being Indonesian citizens aged ≥17 years. Before the questionnaire was distributed, we conducted a readability test to ensure that all the items on the questionnaire could be understood by the respondents in terms of writing and sentence meaning. The readability test was conducted on 6 respondents from September 25, 2021, to September 28, 2021. After the readability test, we made a revision based on the input given by the respondents. Questionnaires were distributed on the web through messaging applications such as WhatsApp (Meta Platforms), Telegram (Telegram FZ LLC), and Line (Line Corporation), as well as social media such as Facebook and Twitter, from October 28, 2021, to November 20, 2021. Questionnaire data were analyzed using descriptive statistics in Microsoft Excel (Microsoft Corp) and SPSS Statistics (version 28; IBM Corp).

To evaluate the artifacts, this study used the evaluation guidelines defined by Venable et al [28] and the evaluation criteria defined by Hevner et al [18]. The goal of the evaluation was to determine the suitability of the design for the needs of health services in Indonesia. The architectural design and application prototype were evaluated qualitatively through interviews with IT or eHealth experts. The evaluation of the architecture design aimed to assess the completeness and conformity with the health system in Indonesia. The evaluation of the application aimed to assess functionality and usability.

This study used the COREQ (Consolidated Criteria for Reporting Qualitative Research) guidelines as a comprehensive checklist that covers the necessary components of qualitative research (Multimedia Appendix 2 [29]). Interview data were analyzed using content analysis techniques in NVivo (version 12; QSR International). The content analysis steps comprised decontextualization, recontextualization, categorization, and compilation [30]. In decontextualization, the authors read the transcribed text and broke down the text into smaller meaning units. Each identified meaning unit was labeled with a code. In recontextualization, the original text was reread alongside the final list of meaning units. In the categorization process, themes and categories were identified. The categorization for the requirement analysis was based on the functionalities and issues in the implementation of PHRs defined by Harahap et al [5]. The categorization for the architectural evaluation interview was carried out based on the architectural components, whereas the categorization for the prototype evaluation interview was carried out based on the implemented functions in the PHR application. At the compilation stage, the authors wrote the results of the analysis.

Ethics Approval

The authors obtained a letter of ethics approval from the Faculty of Computer Science, University of Indonesia, to conduct data collection with letter S-1122A/UN2.F11.D1/PDP.01/2020. The author submitted the letter to the respondents and provided a brief explanation of the study objective. Each respondent verbally provided consent to participate during the interview.

Results

Artifact Development

Respondent Demographics

We interviewed a total of 37 respondents. The respondents were from 10 first-level health facilities (n=6, 60% Puskesmas and n=4, 40% clinics) and 15 referral-level health facilities (n=9, 60% government hospitals and n=6, 40% private hospitals); 5%

(2/37) of respondents were from the Ministry of Health, 3% (1/37) of respondents were from BPJS Kesehatan, and 8% (3/37) of respondents were health application vendors. A total of 73% (27/37) of respondents were male, and 27% (10/37) were female. Most respondents (24/37, 65%) were from the Special Capital Region of Jakarta, whereas others were from Jawa Barat (6/37, 16%), Bali (3/37, 8%), Banten (1/37, 3%), the Special Region of Yogyakarta (1/37, 3%), Riau (1/37, 3%), and Sulawesi Selatan (1/37, 3%). Detailed respondent information is provided in Multimedia Appendix 3.

A total of 1343 respondents filled out the questionnaire. However, there were 24.65% (331/1343) of invalid or duplicate data, so the total valid data from filling out the questionnaire were from 75.35% (1012/1343) of respondents. A total of 37.55% (380/1012) of respondents were male, 62.45% (632/1012) were female, and most (606/1012, 59.88%) lived in Greater Jakarta. Most respondents were aged 20 to 30 years (376/1012, 37.15%; Table 1).

Table 1. Demographics of questionnaire respondents (n=1012).

Demographics	Respondents, n (%)
Sex	
Male	380 (37.5)
Female	632 (62.5)
Age (years)	
17-20	207 (20.5)
20-30	376 (37.2)
31-40	148 (14.6)
41-50	134 (13.2)
51-60	125 (12.4)
>60	22 (2.2)
Domicile	
Greater Jakarta	606 (59.9)
Java island other than Greater Jakarta	279 (27.6)
Outside Java island	78 (7.7)
Education level	
Primary school	0 (0)
Junior high school	1 (0.1)
Senior high school	315 (31.1)
Diploma	62 (6.1)
Bachelor’s degree	415 (41)
Master’s degree	173 (17.1)
Doctorate	46 (4.5)
Familiarity with the use of IT	
Excellent	298 (29.4)
Good	519 (51.3)
Okay	191 (18.9)
Bad	2 (0.2)
Very bad	2 (0.2)

Requirements

Health Organization Requirements

The functions recommended by respondents from health organizations included functions related to access to health records for patients, such as viewing diagnostic data, laboratory and examination results, and medical history:

> *They can view data from medical visits, such as lab results, and x-ray results. Then, you can also see the medical history, including the diagnosis.* [Head of IT, General hospital (GH) 3]

Respondents also mentioned the need for a function to view health facility profiles, such as the list of services and the

availability of beds. Other recommended functions were paying for medical expenses, billing, and claiming health insurance:

> *List of services that can be provided because each hospital is different.* [Physician, GH8]
>
> *If there is a bill for the patient, it will be created to be given to the patient.* [Health application vendor, vendor (VDR) 1]

Respondents also suggested the need for a function for patients to manage information related to medication consumption, such as viewing the history of medication that has been consumed, ordering medication, and scheduling medication consumption:

> *Ordering medicine according to a prescription, then delivery of the medicine.* [Health application vendor, VDR3]
>
> *There is a function for medication consumption.* [Head of IT, primary health care (PHC) 8]

Another recommended function was a feature that patients can use to interact or communicate on the web with medical personnel in health facilities. Communication can occur through chat, messages related to health consultations, or video calls:

> *Communication to hospital and telemedicine with video.* [Head of IT, private hospital (PH) 3]

In addition, respondents mentioned the need for features for patients to manage appointments with medical personnel at health facilities. With this feature, patients can register themselves for treatment at health facilities, including choosing a physician:

> *There is an online registration for patients...there must be information on what time they should be treated.* [Head of IT, GH6]

Respondents also recommended a function for patients to access disease-related information and health tips, such as how to maintain a healthy lifestyle. Others suggested a function for patients to manage their health data to support preventive health efforts or disease prevention. In addition, this function could assist in the recovery from certain diseases that require ongoing health management activities. For example, patients could input data on vital signs and physical activities:

> *There are health articles that can be used when other features are not used. The available information depends on patients' health conditions. This could include articles on healthy lifestyles such as safe cosmetics, or nail care.* [Head of IT planning strategy, health regulator (HR) 2]
>
> *Monitoring tracking is also a very good opportunity because health is not only about curing, we also need preventive measures so that we don't get sick.* [Head of IT, PH2]

In addition to functional requirements, respondents mentioned the need for PHR integration. PHR functionality can be integrated with health applications or other existing data sources, such as electronic medical records in health facilities, to obtain medical summaries, such as diagnoses, laboratory and examination results, and medical history. PHRs also need to be integrated with the referral information system (*sistem informasi rujukan terintegrasi*) to obtain patient referral history and integration with vaccine data, especially for needs during the COVID-19 pandemic:

> *It needs to be integrated with electronic medical record data from health facilities.* [Member of data and information center, HR1]
>
> *We need to integrate data from SISRUTE because it records data from several health facilities. If the patient is referred, the data should be recorded.* [Head of IT, GH5]

PHRs also need to be integrated with BPJS Kesehatan for JKN patients and integrated with web-based payments for patients who seek treatment at health facilities and do not use health insurance:

> *Need to be integrated with BPJS health.* [Head of IT, GH3]
>
> *For payments, it is integrated with online payments.* [Health application vendor, VDR3]

PHRs can be integrated with teleconsultation applications for communication between patients and medical personnel, as well as with pharmacies for ease of ordering medicines. In addition, for the convenience of monitoring personal health data such as physical activities, PHRs need to be integrated with wearable devices. Moreover, PHRs need to be integrated with the healthy family or *Program* Indonesia *Sehat dengan Pendekatan Keluarga* application to support healthy family programs at the Puskesmas and with the national health data repository owned by the Ministry of Health:

> *For prescription, the application needs to be connected to the pharmacy.* [Physician, PH5]
>
> *Regarding fitness, it is difficult to implement, unless you can access data from wearable devices such as smartwatches.* [Member of data and information center, HR1]
>
> *We have a healthy family application (PIS-PK) to record family information related to individual family members, and it is not from the results of the medical examination.* [Head of IT, HR1]

Respondents also mentioned security aspects that need to be applied to PHRs, such as access control, audit trails, data encryption, and data backup. Authentication and authorization are needed in the implementation of PHRs. An audit trail is required to review who is accessing and what data have been accessed in the PHR. A data backup option is required to avoid the risk of data loss. In addition, PHRs need to implement important data encryption, such as passwords:

> *There must be a data backup.* [Head of IT, GH2]
>
> *There is a log in the application to see what time the user logged in and what features were accessed.* [Head of IT, PHC1]

PHRs need to implement user manuals or guidance options to help users understand the information contained in them. In addition, PHRs need to have customization options based on the availability of the internet network as several regions in Indonesia have poor internet connections:

Provide user manuals, video manuals, or readable manuals. [Health application vendor, VDR1]

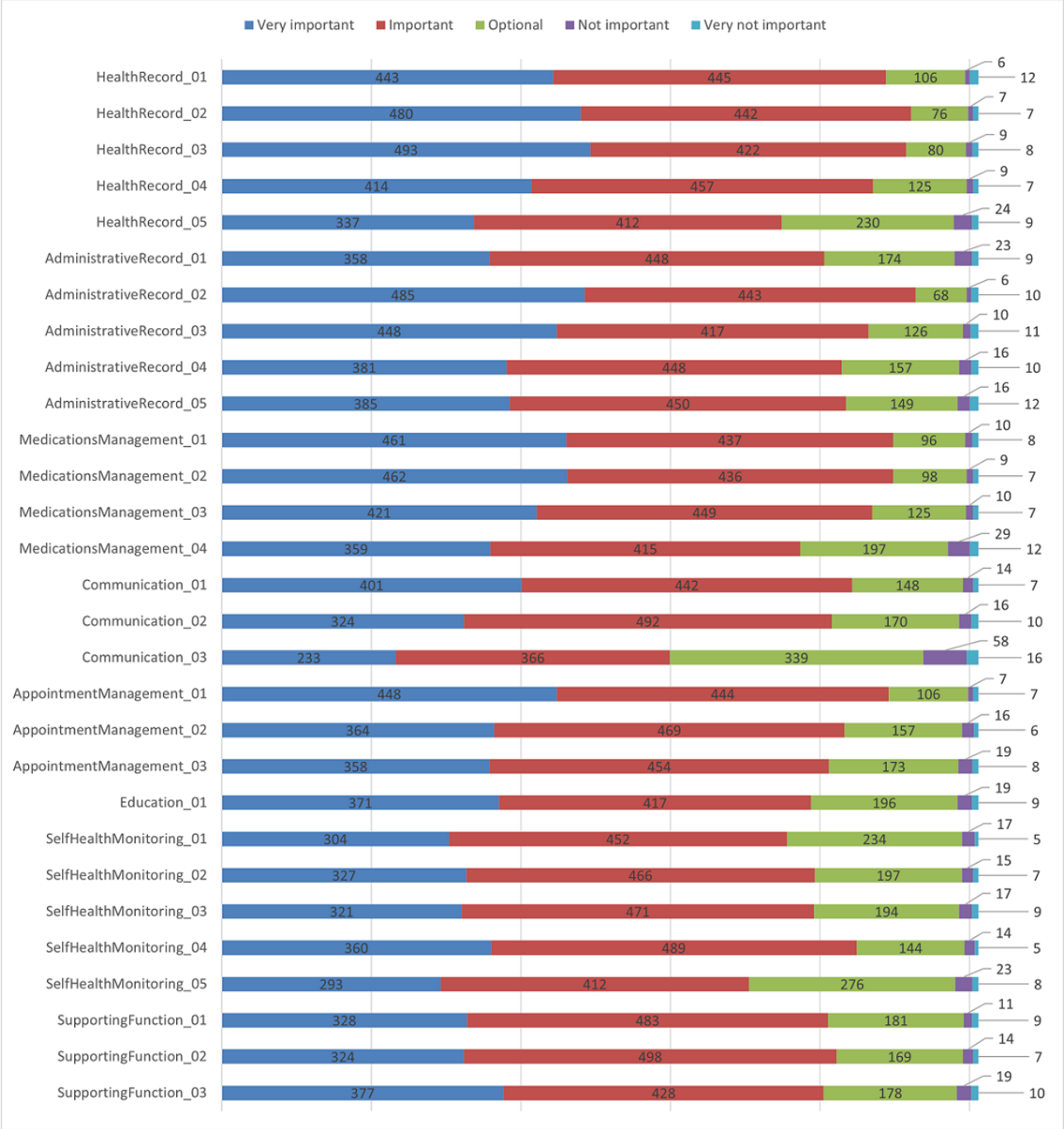
For areas with poor internet network, the application should still be accessible. [Physician, GH8]

Patient Requirements

A total of 70.06% (709/1012) of respondents had used health applications, whereas 29.94% (303/1012) had never used health applications. For each respondent who had used a health application, the questions asked were the health application used, the platform used to access the health application, the length of use of the health application, the frequency of use of the health application in the last 6 months, the features used in the health application, reasons for using health applications, challenges when using health applications, organizations that must be integrated or connected with health applications, and the most important components of a health application (Multimedia Appendix 4).

Respondents were also asked to rate how important the PHR functionalities were based on the previous review (Harahap et al [5]), which comprised health records, administrative records, medication management, communication, appointment management, education, self–health monitoring, and supporting function. The functionality codes for each PHR module are summarized in Multimedia Appendix 5. Respondents were asked to provide an assessment with the following scores: 1=*Very Not Important*, 2=*Not Important*, 3=*Optional*, 4=*Important*, and 5=*Very Important*. The mean was then calculated for each functionality. If the mean of the functionality was <4, then the functionality did not need to be implemented in the integrated PHR model. On the basis of the results of the questionnaire, 27 functionalities had a mean of >4. However, the functionality of sending messages or chatting with support groups and family members in the PHR application (Communication_03) and the functionality to connect wearable devices with the PHR application were considered important by fewer respondents (SelfHealthMonitoring_05), with a mean of <4. Figure 2 shows the respondents' scores for each PHR functionality.

Figure 2. Respondents' scores for each personal health record functionality.



Summary of User Requirements

The requirements of health organizations and patients were grouped into PHR modules and functionalities based on the study by Harahap et al [5], which comprised health records, administrative records, medication management, communication, appointment management, education, and self–health monitoring. In addition, there were emergency modules obtained from patients' requirements as well as security modules and supporting functions to meet the nonfunctional requirements of PHRs. In the health record module, there were functions to view the results of medical examinations, medical history, referrals, and vaccinations. In the administrative records module, there were functions for patient profiles, health facility profiles, physician profiles, health insurance, and payments and billing. In the medication management module, there were functions for medication history, medication reminders, and medication orders. In the communication module, there was a messaging function (SMS text message or video call). In the appointment management module, there were functions for registration, appointment history, reminders or notifications, and ambulance. In the education module, there was a health article function. In the self–health monitoring module, there were health data tracking functions, health dashboards, health calculators, and early warning notifications. In the emergency module, there was an emergency contact function. In the security module, there were authentication, authorization, audit log, and backup functions. In the supporting function module, there were user manual and offline functionalities. A summary of the PHR

module and functionality based on the requirements of the respondent group is provided in Multimedia Appendix 6.

Architecture Development

Overview

This study used The Open Group Architecture Framework (TOGAF) to design the architecture of an integrated PHR system in Indonesia. On the basis of previous studies, TOGAF provides a complete process and methodology to develop architecture [31]. Moreover, TOGAF is the most suitable architectural framework for application in the health sector as it provides a complete architectural development process and can be adapted to the health sector [32,33]. The TOGAF referred to in this study is the TOGAF version 9.2. The Architecture Development Method in the TOGAF can be modified to suit specific needs [34]. In this study, the scope of architectural development was to design information system architecture. Therefore, the TOGAF components needed comprised architecture vision, business architecture, application architecture, data architecture, and technology architecture as required components in designing information system architecture [35].

Architecture Vision

According to the TOGAF 9.2, an architectural vision is a brief description of the target architecture that describes the business value and changes that will result from successful implementation. Architectural vision serves as a vision and boundary in the development of a more detailed architecture [34]. The value that is expected to be provided by the integrated PHR is a complete medical history by allowing patients to obtain medical information from different health facilities. Integrated PHRs can also minimize unnecessary health examinations as patients can share their medical history with their physicians so that physicians have information about previous examinations that have been carried out by patients. In addition, the integrated PHR facilitates communication between patients and physicians and helps patients with administrative activities such as registration, appointments, and payment of medical expenses. Integrated PHRs can also help patients manage health outside the health care environment, such as tracking food consumption and physical activity according to the patient's needs and health conditions.

On the basis of the user requirements and literature review, we formulated architecture principles for integrated PHRs in Indonesia following the TOGAF 9.2 (Table 2). These comprise business, data, application, and technology principles [34]. The business principles comprise information management as everybody's business, business continuity, service orientation, compliance with the law, and patient-centeredness. Data principles comprise data being an asset, shared, and accessible; common vocabulary and data definitions; and data security. Application principles comprise technological independence and ease of use as well as functionality completeness. The technology principles comprise interoperability and ease of access.

Table 2. Architecture principles.

Domain and principle	Description	References
Business		
Information management is everybody's business	To support health services, health organizations and patients need to be involved in managing information on the PHR[a].	• The Open Group [34] • Harahap et al [36]
Business continuity	The PHR has an optional function that allows users to use it with a poor internet connection.	• The Open Group [34] • Harahap et al [36]
Service orientation	The services provided by the PHR are integrated with health care activities in Indonesia.	• The Open Group [34] • Harahap et al [36]
Compliance with the law	The PHR needs to comply with all applicable laws, policies, and regulations in Indonesia.	• The Open Group [34] • Harahap et al [36]
Patient-centeredness	The PHR is designed for patients to have health information from various sources.	• The Open Group [34] • Harahap et al [36]
Data		
Data are an asset	The data on the PHR could assist the decision-making of health care providers or patients in managing their health.	• The Open Group [34] • Harahap et al [36]
Data are shared	Data can be shared between patients and health care providers.	• The Open Group [34] • Harahap et al [36]
Data are accessible	Access to accurate data is needed to improve quality and efficiency in the management of patient health.	• The Open Group [34] • Harahap et al [36]
Common vocabulary and data definition	The data on the PHR must have the same definition across health organizations to allow for data sharing.	• The Open Group [34] • Harahap et al [36]
Data security	A security mechanism is needed to protect the data stored or exchanged on the PHR.	• Harahap et al [5] • The Open Group [34] • Harahap et al [36]
Application		
Technology independence	The PHR can be integrated with various applications on different platforms.	• The Open Group [34] • Harahap et al [36]
Functional completeness	The PHR provides functionalities to support promotive, preventive, curative, and rehabilitative health services in Indonesia.	• Harahap et al [5,36]
Ease of use	The PHR should be easy to use so that users can complete their tasks.	• Harahap et al [5] • The Open Group [34] • Harahap et al [36]
Technology		
Interoperability	The PHR needs to have interoperability standards for sharing data among stakeholders and health information systems.	• Harahap et al [5] • The Open Group [34] • Harahap et al [36]
Ease of access	The PHR needs to be implemented using common technology used by the community to facilitate easy access to information.	• Harahap et al [36] • Kharrazi et al [37]

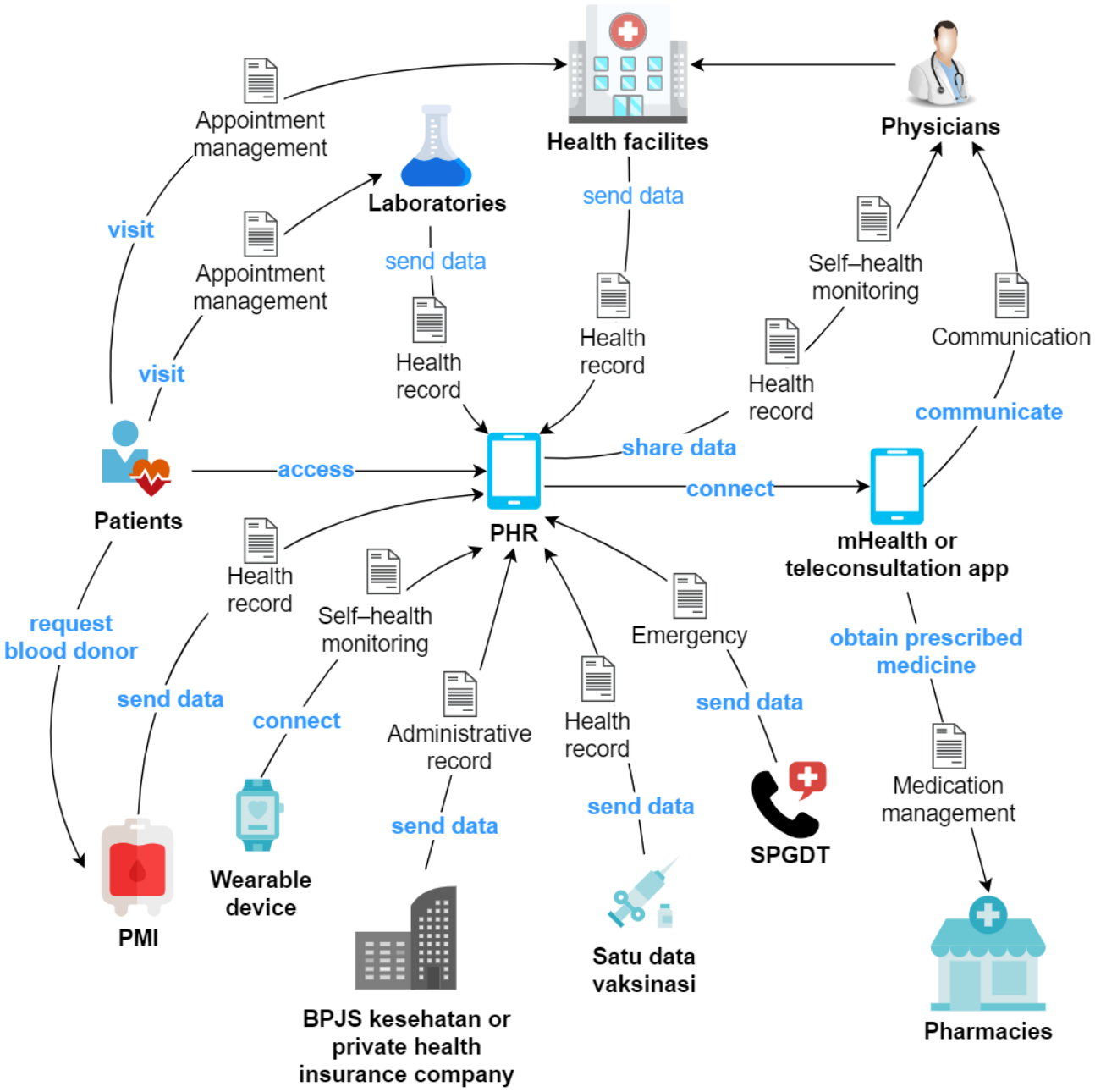
[a]PHR: personal health record.

Business Architecture

Business architecture defines the business strategy, governance, organization, and key business processes [34]. On the basis of the interviews with participants from health facilities and a review of health regulations in Indonesia, the business process for PHRs can be divided into health care business processes, self–health monitoring business processes, vaccination business processes, and home care business processes. Each of these process flows can be seen in Multimedia Appendix 7. The main parties involved in the PHR are patients, physicians, health facilities, laboratories, the *Palang Merah Indonesia* or Indonesian Red Cross (PMI), BPJS Kesehatan or other private health insurance companies, and pharmacies.

On the basis of the flow of the process, we added a rich picture to describe the data exchange on PHRs in the health care process (Figure 3). Patients register to make visits to health facilities or laboratories. Patients who need blood donors can also submit a blood donation request to the PMI. Patients can share their health records or personal health monitoring data with their physicians. Patients who have received health services will receive their data in the PHR. Patients can connect the PHR with wearable devices to record personal health monitoring data. Patients can connect the PHR to BPJS Kesehatan or other health insurance companies to obtain their status. Patients can communicate with their physicians or receive their prescribed medicine at pharmacies through the mHealth apps or teleconsultation applications that are connected with the PHR. Moreover, patients can receive vaccination data through the PHR, which is linked to the Ministry of Health vaccine data (*satu data vaksinasi*). The PHR can also connect with an integrated emergency management system (*sistem penanggulangan gawat darurat terpadu*) to obtain the nearest emergency service contacts.

Figure 3. Rich picture of personal health record (PHR) data exchange in the health care process. BPJS: Badan Pelaksana Jaminan Sosial Kesehatan or Social Security Agency for Health; mHealth: mobile health; PMI: Palang Merah Indonesia or Indonesian Red Cross; SPGDT: sistem penanggulangan gawat darurat terpadu or integrated emergency management system.

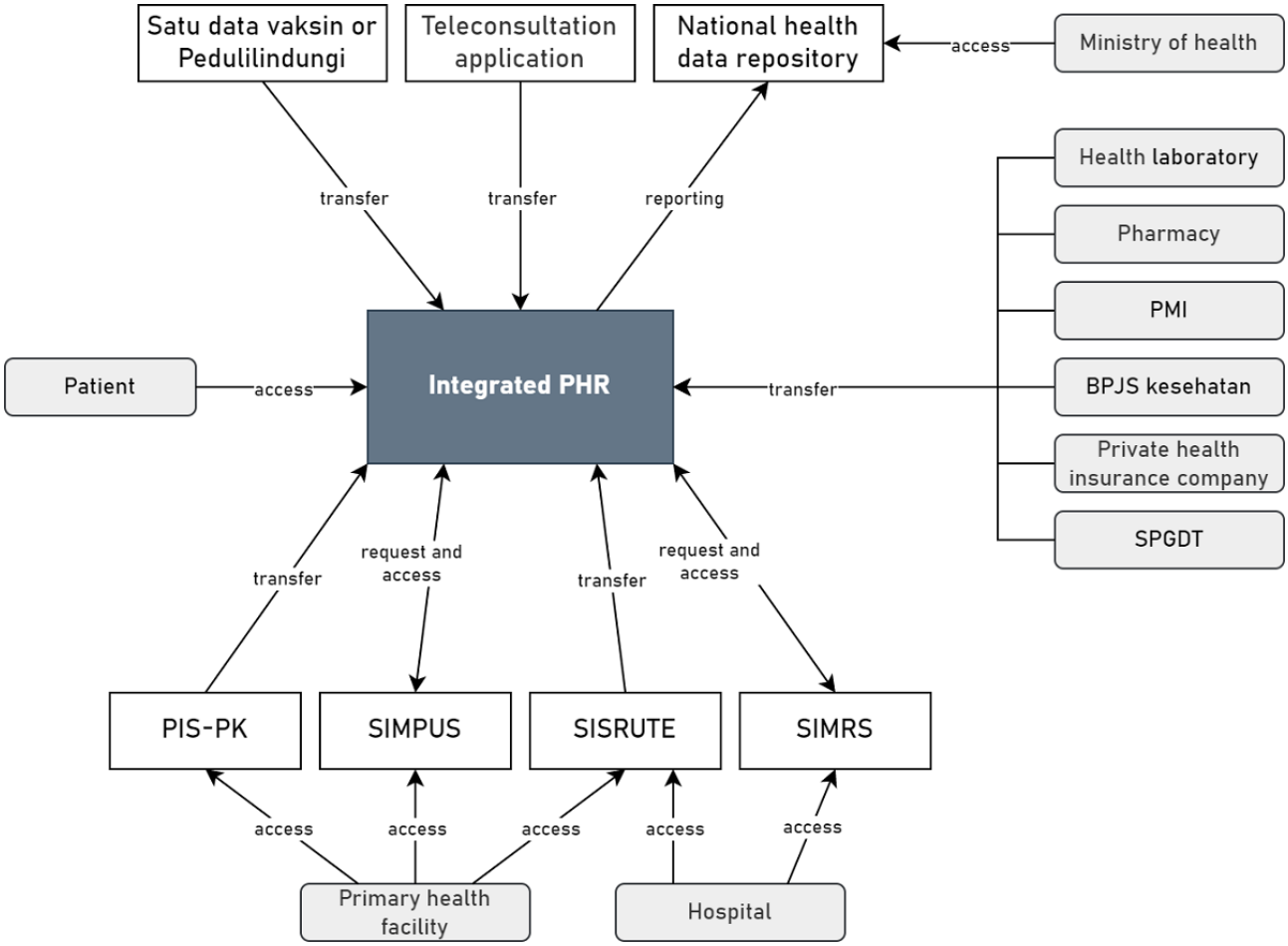


Application Architecture

The TOGAF 9.2 defines application architecture as a blueprint for the applications to be developed, their interactions, and their relationship to the organization's core business processes [34]. On the basis of the results of interviews and questionnaires, parties that need to be integrated into PHRs are primary health facilities (Puskesmas and clinics), referral health facilities (government and private hospitals), health laboratories,

pharmacies, *satu data vaksinasi*, mHealth app or teleconsultation application providers, BPJS Kesehatan, private health insurance companies, and the Ministry of Health. In addition, PHRs need to have the option to connect with wearable devices for tracking health data and family members. On the basis of the results of the business architecture design, other parties that need to be integrated with PHRs are PMI and the integrated emergency management system (*sistem penanggulangan gawat darurat terpadu*). Health facilities are the parties that are most integrated with the information in the PHR. The PHR is integrated with hospital information systems (*sistem informasi manajemen rumah sakit*) in referral health facilities and primary health care information systems (*sistem informasi puskesmas*) in primary health facilities to access health records. In addition to hospital information systems (*sistem informasi manajemen rumah sakit*) and primary health care information systems (*sistem informasi puskesmas*), the PHR is integrated with an integrated referral information system (*sistem informasi rujukan terintegrasi*), which is used by health facilities in Indonesia to access patient referrals. Especially for Puskesmas, the PHR can be integrated with the *Program* Indonesia *Sehat dengan Pendekatan Keluarga* information system. Figure 4 summarizes the integration of the PHR with health information systems or health care providers in Indonesia.

Figure 4. Data exchange of the personal health record (PHR) in Indonesia. BPJS: Badan Pelaksana Jaminan Sosial Kesehatan or Social Security Agency for Health; PIS-PK: Program Indonesia Sehat dengan Pendekatan Keluarga; PMI: Palang Merah Indonesia or Indonesian Red Cross; SIMPUS: sistem informasi puskesmas or primary health care information system; SIMRS: sistem informasi manajemen rumah sakit or hospital information system; SISRUTE: sistem informasi rujukan terintegrasi or referral information system; SPGDT: sistem penanggulangan gawat darurat terpadu or integrated emergency management system.

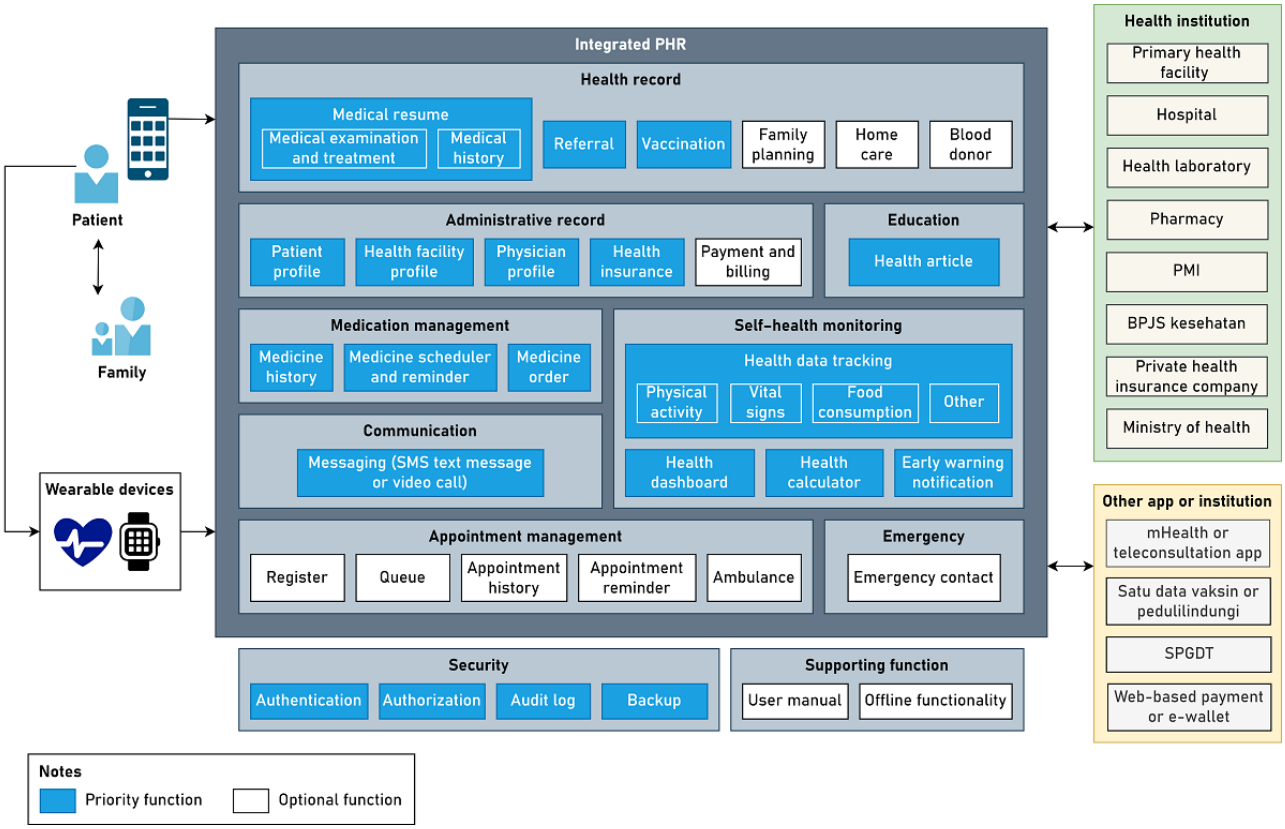


The modules in the PHR comprise health records, administrative records, medication management, communication, appointment management, education, self–health monitoring, security, supporting functions, and emergency. In the health record module, the functionalities comprise a medical summary (results of physical examinations and medical support as well as disease and medication history), referrals, and vaccinations. This module also adds the functions of family planning, home care, and blood donors based on the identification of activities in the business architecture. In the administrative record module, the functionalities comprise patient profiles, health facility profiles, health personnel profiles, health insurance, and payments and billing. In the medication management module, the functionalities comprise medication history, medication reminders, and medication orders. In the communication module, the functionalities comprise messaging. In the appointment management module, the functionalities comprise registration, appointment history, reminders, notifications for appointments, and ambulance services. In the education module, the functionalities comprise health articles containing information on disease problems and health tips. In the self–health monitoring module, the functionalities comprise health data tracking, health dashboards, health calculators, and

alert notifications. In the emergency module, the functionalities comprise emergency contacts.

We also designed the modules and functionalities to meet the nonfunctional requirements of PHRs. The modules comprise security and supporting functions. In the security module, the functionalities comprise authentication, authorization, audit logs, and data backup. In the supporting function module, the functionalities comprise user manual and offline functionality. The functionalities that need to be prioritized for implementation in PHRs are functions related to health management (health care, health prevention, and health promotion) and functions to support information security, whereas other functionalities such as functions related to administration and supporting functions are optional. To access the PHR, the platform used is a smartphone as, based on the results of the questionnaire, it is the most widely used platform by the public to access health applications. Figure 5 summarizes the integrated PHR system model in Indonesia.

Figure 5. Modules and functionalities of the integrated personal health record (PHR) system in Indonesia. BPJS: Badan Pelaksana Jaminan Sosial Kesehatan or Social Security Agency for Health; mHealth: mobile health; PMI: Palang Merah Indonesia or Indonesian Red Cross; SPGDT: sistem penanggulangan gawat darurat terpadu or integrated emergency management system.



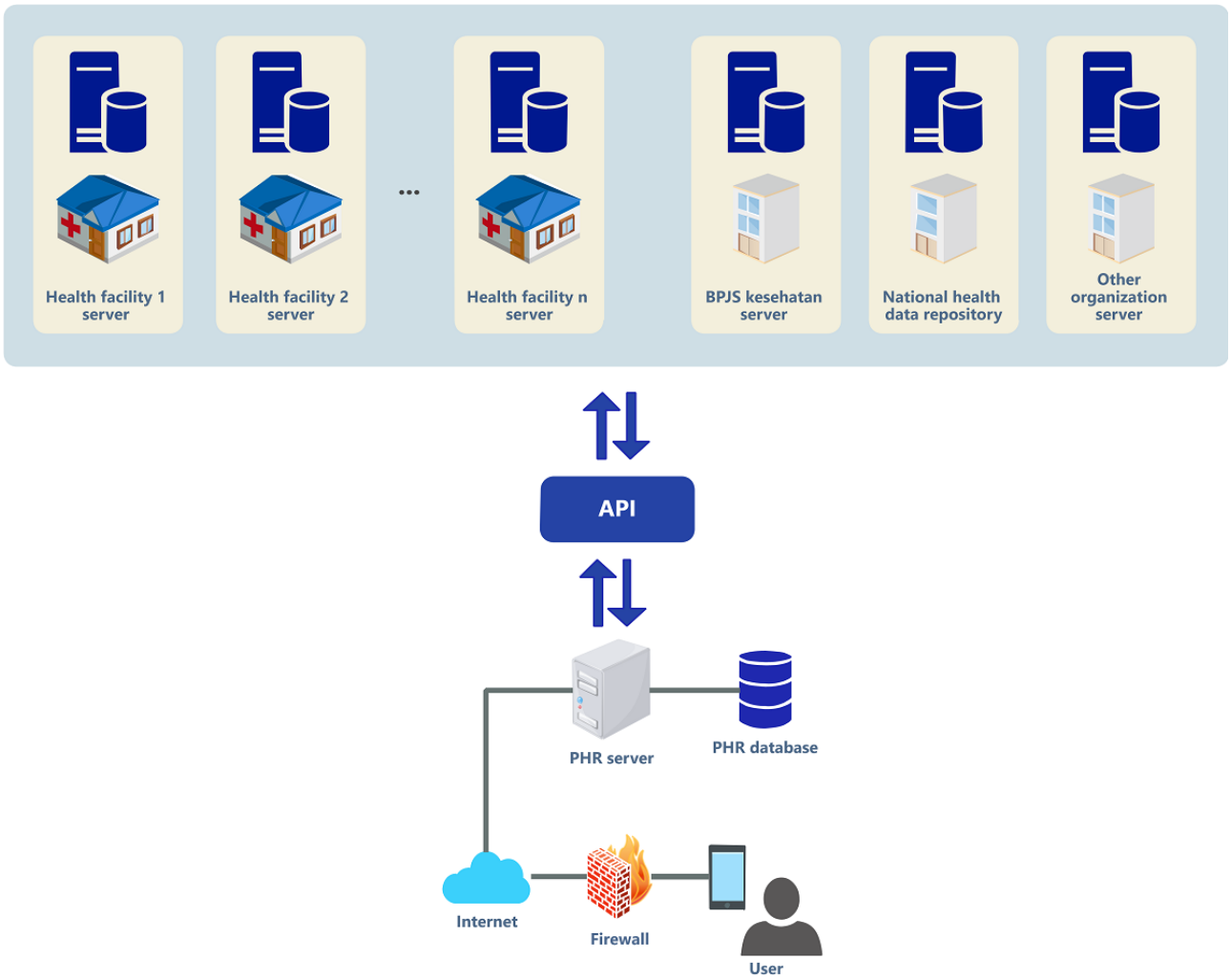
Data Architecture

The TOGAF 9.2 defines data architecture as the logical and physical structure of the organization's data assets and data management resources [34]. We grouped data in PHRs into 3 data categories comprising master, transaction, and reference data [38,39]. Transaction data relate to data that are recorded every time a transaction occurs, such as medical records, vaccinations, and health referrals. Master data are data that do not change and do not need to be recorded in every transaction, such as patient and health facility data. Reference data are a collection of values or classifications that can be referenced by master and transaction data.

Technology Architecture

The technology architecture for PHRs is described in the form of a high-level architecture to illustrate the technology required for PHR implementation and integration with other systems (Figure 6). Patients or individuals access the PHR through a mobile app. PHR development uses the React Native cross-platform app development that can be implemented on the Android or iOS platforms. The PHR application is accessed by users via the internet, and a firewall is used for security. The PHR server comprises an application server and a database server. The application server provides access to data for the user, whereas the database server provides the data requested by the application server [40].

Figure 6. Technology architecture. API: application programming interface; BPJS: Badan Pelaksana Jaminan Sosial Kesehatan or Social Security Agency for Health; PHR: personal health record.



An application programming interface (API) is used as an intermediary for interaction between the PHR and other information systems. The type of API used is Fast Healthcare Interoperability Resources (FHIR). FHIR is an international standard recommended by the Ministry of Health to solve the problem of data exchange in health information systems in Indonesia [17]. FHIR is flexible and can be adapted to stakeholder needs, clinical specifications, and health policies. FHIR can be used to manage a single data entity (eg, heart rate), groups of data entities (eg, vital signs, medications, and allergies), or electronic recording systems such as PHRs. Therefore, FHIR is suitable for exchanging data on PHRs as PHRs aim to collect and exchange individual health data [41].

Prototype Development

The main actors involved in health care and management activities are patients and physicians. Patients are actors who receive health services and play a role in managing their health through applications. Physicians are actors who provide health services to patients. Physicians comprise general practitioners and specialists. The functionalities developed in the prototype design are priority functions defined in the application architecture, which comprise medical summaries, referrals, vaccinations, health facility profiles, physician profiles, patient profiles, messaging, medication history, medication reminders, medication orders, health data tracking, health calculators, health articles, and notifications. An explanation of the design requirements for each function and the actors involved is presented in Table 3. We developed a high-fidelity prototype for a mobile app. Some examples of patient and physician prototype designs are shown in Figures 7 and 8, respectively.

Table 3. Design requirements for application prototype.

Function	Description	Actor
Medical summary	View the history of patient visits to health facilities, including detailed examination results	Patient and physician
Referral	View patient referral history and detailed examination results from each patient referral	Patient and physician
Vaccination	View a patient's vaccination history	Patient and physician
Health facility profile	Search for health facilities and see the nearest health facility	Patient
Health facility profile	View detailed health facility information	Patient and physician
Physician profile	View the profile of physicians who have treated patients	Patient
Physician profile	Manage physician profile	Physician
Patient profile	Manage patient profiles, including adding family members, connecting with BPJS[a] Kesehatan or other health insurance companies, and connecting with wearable devices	Patient
Patient profile	View a list of patients who have been treated	Physician
Messaging	View lists and details of messages between patient and physician	Patient and physician
Medication history	View a list of past or current medications	Patient
Medication reminder	Manage reminders to take medication	Patient
Medication order	Obtain the prescribed medicine at the pharmacy	Patient
Health data tracking	Record health monitoring data such as physical activity, food consumption, and others	Patient
Health data tracking	View the health monitoring dashboard that has been created by the patient	Patient and physician
Health calculator	Perform patient health calculations such as BMI	Patient
Health article	Read health articles, such as information about disease problems or health tips	Patient
Notification	Receive notifications such as incoming messages, documents sent, or reminders to perform certain activities	Patient and physician

[a]BPJS: Social Security Agency for Health or Badan Pelaksana Jaminan Sosial Kesehatan.

Figure 7. Example of the home page, medical summary, and health data tracking in the patient prototype.

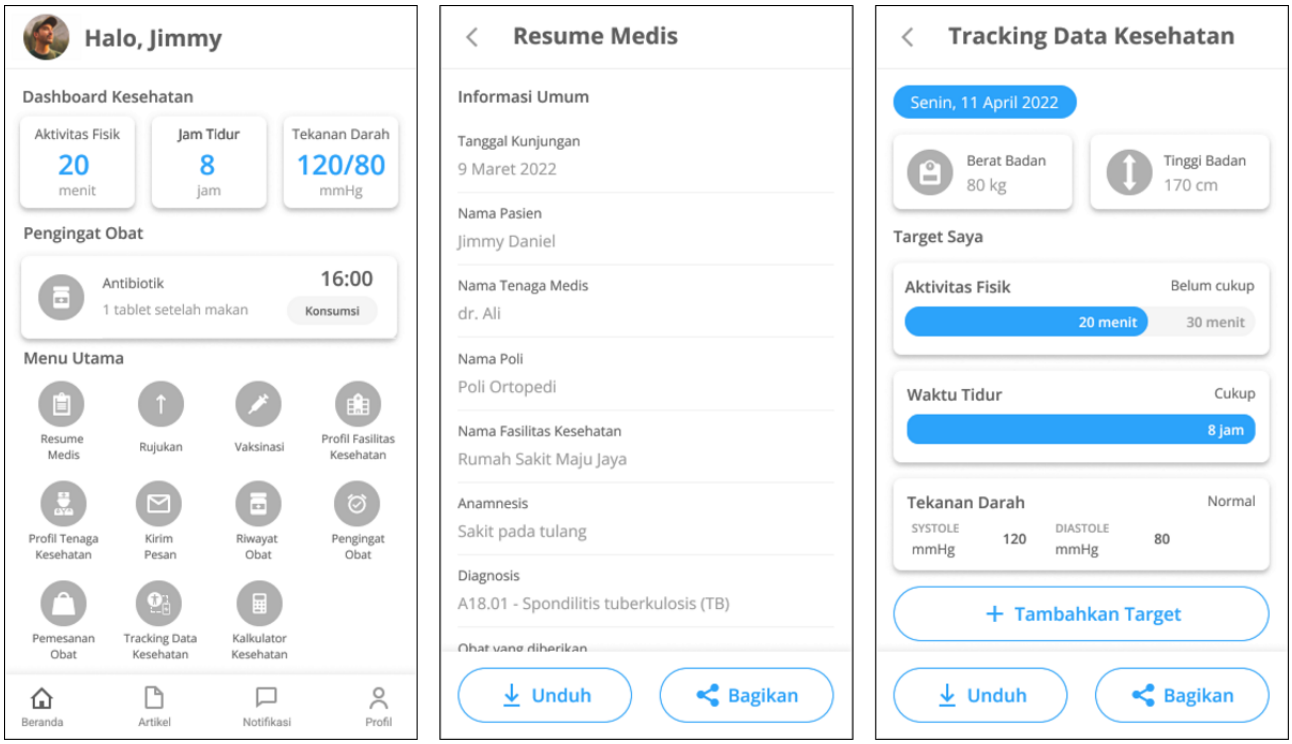
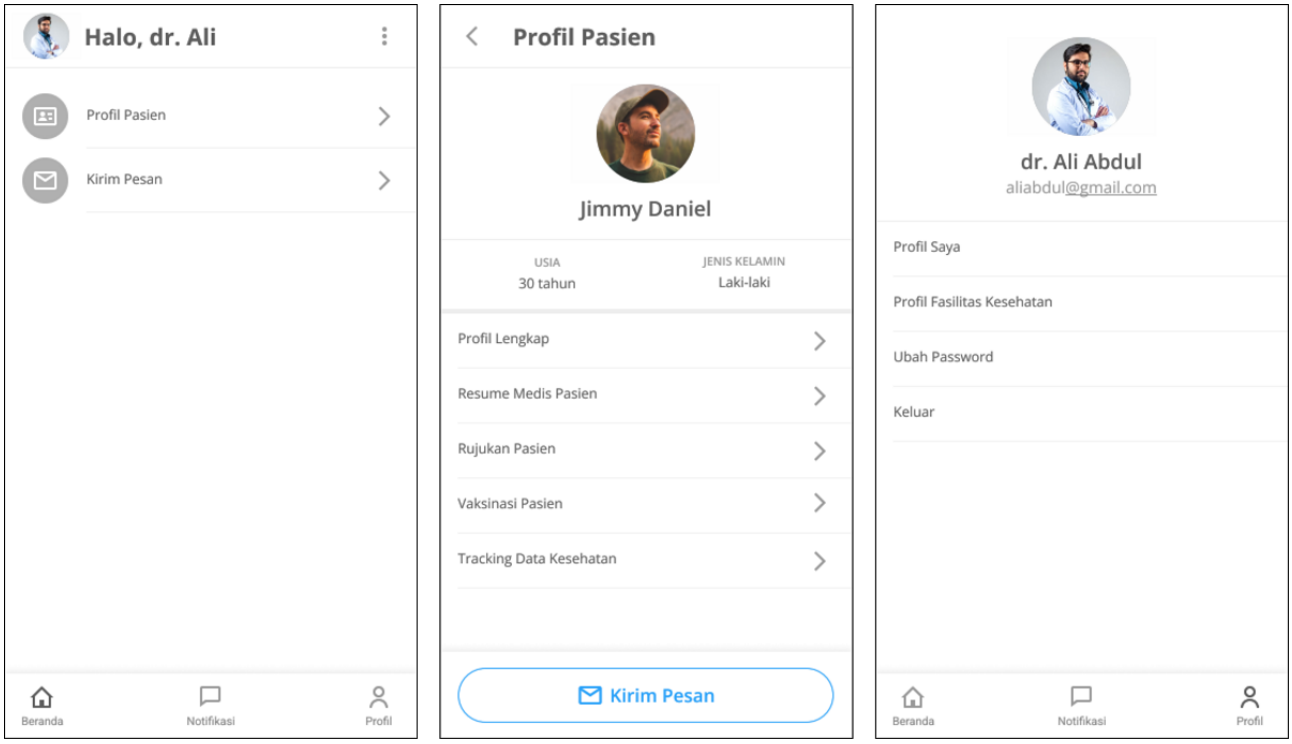


Figure 8. Example of the home page, patient profile, and physician profile from the physician prototype.



Artifact Evaluation

Respondent Demographics

To evaluate the architectural design and application prototype, we conducted interviews with IT or eHealth experts. Interviews were conducted with 6 respondents: 1 (17%) respondent from the Ministry of Health, 1 (17%) academician, 1 (17%) respondent from a government hospital, 1 (17%) respondent from a private hospital, and 2 (33%) health application vendors. Interviews were conducted from April 5, 2022, to April 8, 2022, with an interview duration of 40 to 60 minutes. The information of the respondents is presented in Table 4.

Table 4. Respondent demographics.

Respondent code	Sex	Role	Work experience (years)
E1	Male	Health application vendor	1-5
E2	Female	Academician	1-5
E3	Male	IT management in the Ministry of Health	>10
E4	Male	IT management in a government hospital	>10
E5	Male	IT management in a private hospital	>10
E6	Male	Health application vendor	>10

Architecture Evaluation

The evaluation was carried out to assess the suitability of the integrated PHR design for the needs of health services in Indonesia (Table 5). All respondents (6/6, 100%) stated that the architectural vision and business architecture in the PHR architectural design described the needs of health services in Indonesia. Regarding the integration in the application architecture, there were several recommendations from respondents regarding parties that needed to be integrated with the PHR but were not described in the application architecture design. Other parties that need to be integrated with the PHR include billing gateways for the payment function (respondent E1) and the Directorate General of Population and Civil Registration for patient identity (respondents E2 and E3). A respondent (E4) suggested adding a health screening function to the PHR. Another respondent (E3) suggested making the messaging and medication order functions optional as they can be connected with other applications. Regarding data architecture, respondent E6 commented that the data architecture was sufficient as long as there was an explanation of the data source in the PHR. For security, respondents suggested more options for authentication methods, such as biometrics (respondent E1) and face recognition (respondent E5). Regarding technology architecture, 100% (6/6) of the respondents stated that the use of an API was a suitable solution for integration between PHRs and other applications in Indonesia. Respondent E3 commented that FHIR was the right type of API to use for PHR implementation. For security and privacy needs, respondents also agreed with the use of firewalls in the technology architecture.

On the basis of the evaluation with IT and eHealth experts, improvements were made by adding billing gateways and Directorate General of Population and Civil Registration to the application architecture (Figure 9). Improvements were also made to the modules and functionality of the integrated PHR system in Indonesia in application architecture (Figure 10). In the medication management module, the medication order function was changed from a priority function to an optional function that can be connected with mHealth apps or teleconsultation applications in Indonesia. In the communication module, the messaging function was also changed from a priority function to an optional function that can be connected with mHealth apps or teleconsultation applications in Indonesia. Improvements were also made by adding a health screening function to the self–health monitoring module. In the security module, authentication methods were added, including passwords, biometrics, and face recognition. Descriptions of each module and functionality in the PHR are summarized in Multimedia Appendix 8. Improvements to the data architecture were made to add health screening data to the transaction data category. Data categories with data groups and descriptions in the PHR are described in Multimedia Appendix 9.

Table 5. Summary of personal health record (PHR) architecture evaluation results.

Architecture component	Evaluation criteria	Evaluation results	Respondent
Architecture vision	Suitability for the health services in Indonesia	It is sufficient to describe the needs of health services in Indonesia	E1, E2, E3, E4, E5, and E6
Business architecture	Conformity with the process of health services in Indonesia	It is sufficient to describe the needs of health services in Indonesia	E1, E2, E3, E4, E5, and E6
Application architecture	Integration with health information systems or other parties	Integration with billing gateway	E1
Application architecture	Integration with health information systems or other parties	Integration with Dukcapil[a]	E2 and E3
Application architecture	Completeness of functionality	Messaging and medication orders as optional functions	E3
Application architecture	Completeness of functionality	Addition of health screening function	E4
Application architecture	Completeness of functionality	Addition of authentication method options such as biometrics and face recognition	E1 and E5
Data architecture	Data requirements and completeness	Data architecture is sufficient as long as there is an explanation of the data source in the PHR	E6
Technology architecture	Technology requirements for PHR implementation	The architecture already describes the technology requirements for PHR implementation	E1, E2, E3, E4, E5, and E6

[a]Dukcapil: Directorate General of Population and Civil Registration.

Figure 9. Design improvements to the data exchange in the personal health record (PHR) in Indonesia. BPJS: Badan Pelaksana Jaminan Sosial Kesehatan or Social Security Agency for Health; Dukcapil: Directorate General of Population and Civil Registration; PIS-PK: Program Indonesia Sehat dengan Pendekatan Keluarga; PMI: Palang Merah Indonesia or Indonesian Red Cross; SIMPUS: sistem informasi puskesmas or primary health care information system; SIMRS: sistem informasi manajemen rumah sakit or hospital information system; SISRUTE: sistem informasi rujukan terintegrasi or referral information system; SPGDT: sistem penanggulangan gawat darurat terpadu or integrated emergency management system.

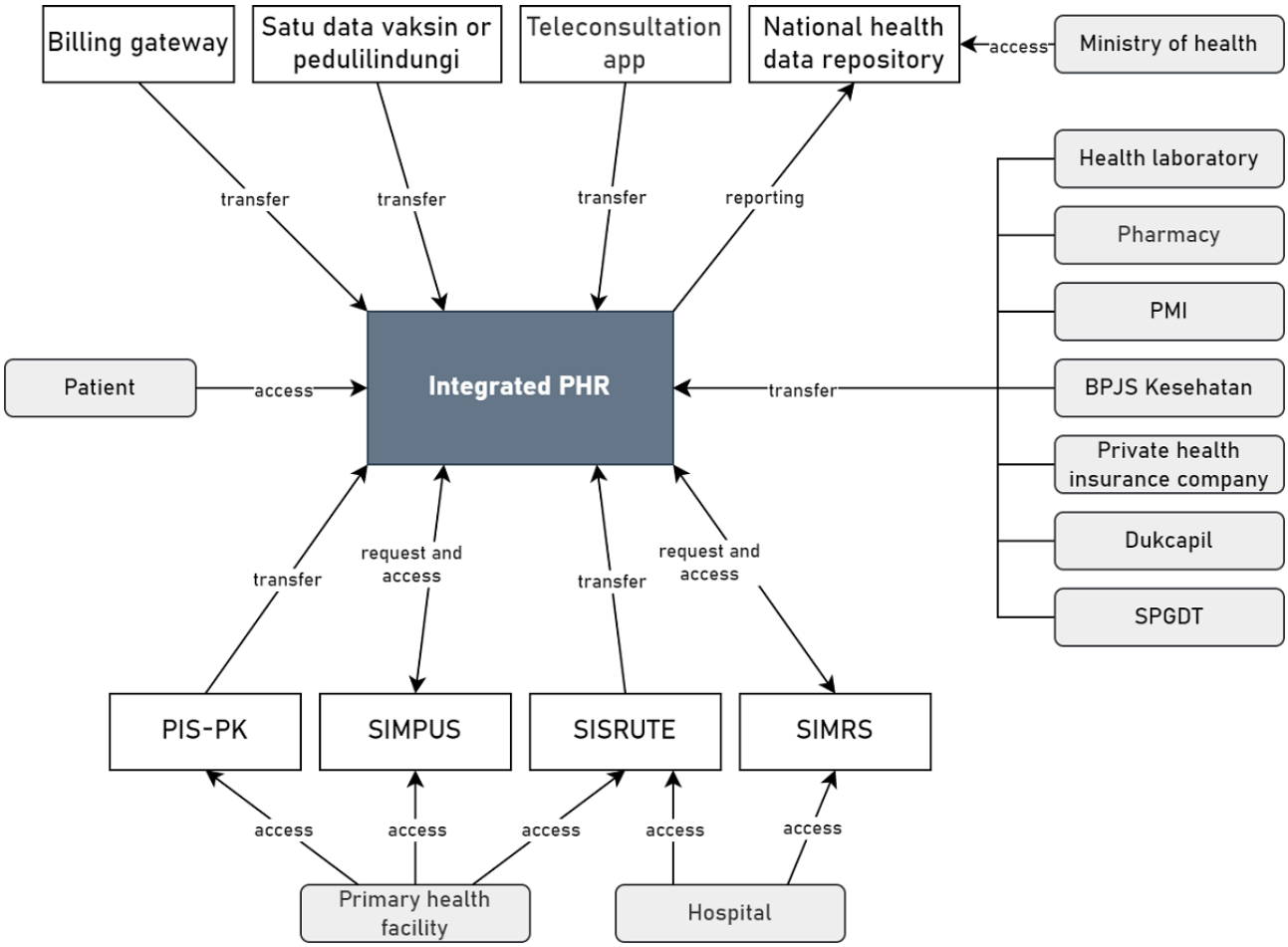
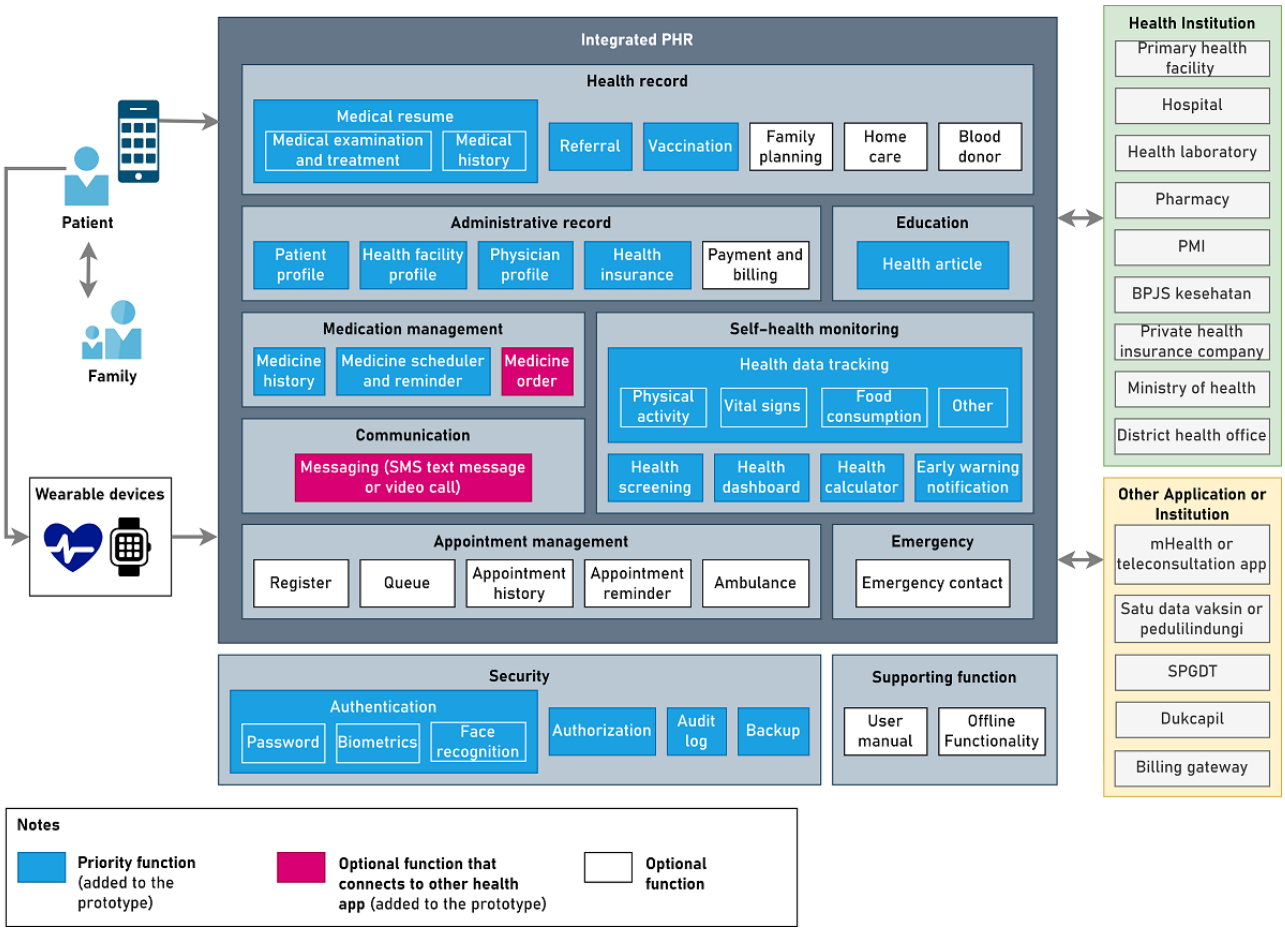


Figure 10. Design improvements to the modules and functionalities of the integrated personal health record (PHR) system in Indonesia. BPJS: Badan Pelaksana Jaminan Sosial Kesehatan or Social Security Agency for Health; Dukcapil: Directorate General of Population and Civil Registration; mHealth: mobile health; PMI: Palang Merah Indonesia or Indonesian Red Cross; SPGDT: sistem penanggulangan gawat darurat terpadu or integrated emergency management system.



Prototype Evaluation

The evaluation of the prototype design resulted in suggestions for improvements related to the main functions developed in the prototype design (Table 6). Suggestions for improvements that need to be made were the functions of medical summary, referral, vaccination, physician profile, messaging, medication history, medication reminder, medication order, health data tracking, notification, and patient profile. In the medical summary, referral, and vaccination functions, a respondent (E3) suggested adding a patient identification number. Some examples of the prototype improvements are shown in Figure 11.

In the medical summary function, suggestions for improvement were the addition of the patient's overall medical history (respondent E3) and the addition of the patient's medical record number (respondent E5) to the medical summary details. The addition of the patient's overall medical history aimed to make it easier for patients to view their complete medical history without having to look at the details of each medical summary one by one. The patient's medical record number was intended to be the patient's identity number at the health facility.

In the referral function, suggestions for improvement were the addition of information on the actions given before the patient was referred (respondent E1) and the reason for the patient being referred (respondent E2) in the referral details. Other suggestions for improvement were the addition of information on the type of referral, such as back referral (respondent E2). This information is needed for the continuous treatment of patients as health workers who treat patients need to know the complete condition of the patient.

In the vaccination function, suggestions for improvement included adding other types of vaccinations apart from COVID-19 (respondents E1, E2, E3, and E4). Additional types of vaccinations are needed so that this function can be used not only during the COVID-19 pandemic but also after it ends. In addition, in the vaccination function, it was recommended that the patient be able to view the vaccination history of family members (respondent E1).

In the physician profile function, the suggestion for improvement was changing the physician's ID to the *Surat Izin Praktik* number (respondents E1, E2, and E4) in the detail of the physician profile as the *Surat Izin Praktik* number is the standard numbering used for every health facility in Indonesia. Another suggestion for improving this function was the deletion of information to view the recommendations of the nearest physician (respondents E3 and E5). This is because, in implementation, it will be difficult to obtain updated information on physicians' availability around the patient's location.

In the messaging function, the suggestions for improvement were to link this function to health applications or teleconsultation applications that already exist in Indonesia. The goal is to facilitate the implementation of the PHR application as other applications can handle the messaging function. This respondent (E3) also suggested doing the same for the medication order function.

In the medication history function, a suggestion for improvement was given by a respondent (E1) to add medication categories, such as generic or patent. The respondent also gave suggestions on the medication reminder function to add information on how to take the medication. Another suggestion for improvement in medication reminders was the need to add information on whether the medication should be fully consumed (respondent E5).

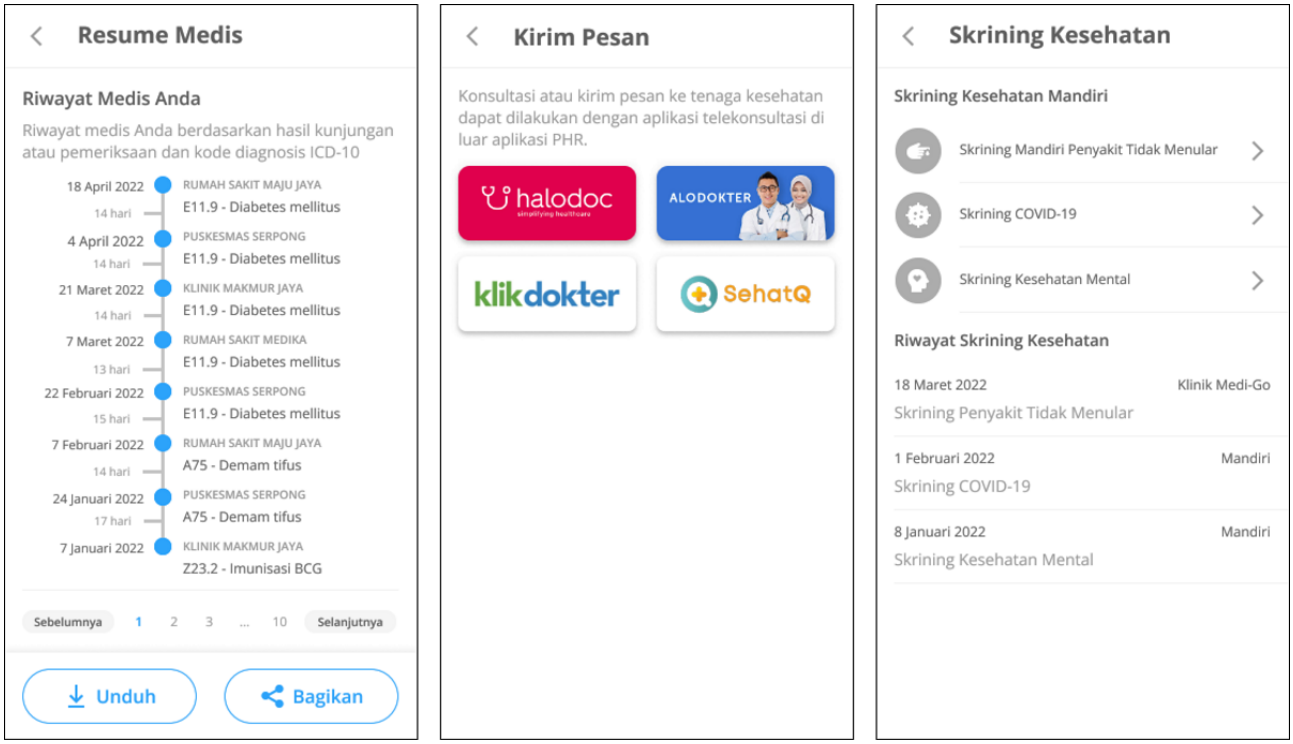
Table 6. Summary of prototype evaluation results.

Actor and function	Evaluation result
Patient	
Medical summary	• Add a medical summary dashboard to view the patient's overall medical history (respondent E3) • Add the patient's NIK[a] number (respondent E3) • Add the patient's medical record number to the medical summary details (respondent E5)
Referral	• Add reason for referral (respondent E2) • Add medication treatment information to referral details (respondent E2) • Add referral type (respondent E2) • Add the patient's NIK number (respondent E3)
Vaccination	• Add vaccine category other than COVID-19 vaccine (respondents E1, E2, E3, and E4) • Add the patient's NIK number (respondent E3)
Physician profile	• Change physician ID to SIP[b] number (respondents E1, E2, and E4) • Delete nearest physician recommendations (respondents E3 and E5)
Messaging	• Link to other health applications (respondent E3)
Medication history	• Add patent or generic medication category (respondent E1)
Medication reminder	• Add the way of taking the medication (respondent E1) • Add information on whether the medication must be fully consumed or not (respondent E5)
Medication order	• Link to other health applications (respondent E3)
Health data tracking	• Add weekly health data tracking dashboard (respondent E1) • Add sample health data tracking input page (respondent E5)
Notification	• Display changes to distinguish read from unread notifications (respondent E1)
Patient profile	• Add other security options such as biometrics (respondent E1) • Add blood type to patient profile details (respondent E2) • Add marital status to patient profile details (respondent E5)
Other	• Add log-in or registration options with Gmail (respondent E1) • Add health screening function (respondent E4)
Physician	
Patient profile	• Add a dashboard to view patients who need to be responded to (respondent E1) • Add menu of patient health screening results (respondent E4)
Notification	• Add sample notification for referral (respondent E2)
Physician profile	• Delete the function to view the health facility profile on the physician profile (respondents E4 and E5)

[a]NIK: Nomor Induk Kependudukan (patient identification number).

[b]SIP: Surat Izin Praktik number (standard numbering used for every health professional or health care provider [physician, nurse, and midwife] in Indonesia).

Figure 11. Examples of prototype improvements.



In the health data tracking function, suggestions for improvement were the need for a weekly health dashboard to observe trends in patient health tracking (respondent E1). Another suggestion was to add a sample data input page for this function (respondent E6). In the notification function, the suggestion for improvement was the need for display adjustments to indicate the difference between read and unread notifications (respondent E1).

In the patient profile function, suggestions for improvement included the need for security options other than passwords, such as biometrics (respondent E1). In the detailed patient profile, additional information was needed, such as blood type (respondent E2) and marital status (respondent E5). In addition to the functions that have been discussed, respondent E4 gave suggestions to add a health screening function to the PHR, including independent health screening or health screening in primary health facilities.

In the evaluation of the prototype design for the physician, suggestions for improvement were the addition of a dashboard to see patients who need to be responded to (respondent E1) and the addition of patient health screening results following the suggestions for the prototype design for patients (respondent E4). Another improvement suggestion was the addition of a notification example for referral data that are sent from patients to the physician (respondent E2). Furthermore, for the user profile function, suggestions for improvement were to remove unnecessary information, such as the profile function of health facilities (respondents E4 and E5).

Discussion

Principal Findings

This study designed an integrated PHR system architecture in Indonesia and an application prototype. In Indonesia, there are various mHealth apps or teleconsultation applications, such as AloDokter, Halodoc, and Mobile JKN. AloDokter and Halodoc are connected to health care providers, mostly private clinics and hospitals, whereas Mobile JKN is for JKN patients in primary health facilities such as Puskesmas and clinics. However, the exchange of health information in these applications is one-way, from the health facility or physician to the patient. Previous research on the adoption of PHRs has also found that health facilities in Indonesia generally do not provide web-based access to patients' health records [36]. This study provides a PHR model that is connected with various health care providers and is integrated into routine health care practice in Indonesia.

A review conducted by Hoque et al [42] stated that most health applications or mHealth research in low- and middle-income countries do not follow the design science approach. The DSR approach informs how artifact validation is carried out so that it can provide evidence that the designed artifact is useful and meets the users' requirements [21]. This study uses a DSR approach with evaluations carried out by IT or eHealth experts so that the PHR design follows the practice of health services in Indonesia.

The architecture development based on the TOGAF can describe the need for integration into the PHR by describing who are the stakeholders involved in the PHR. Architecture development using the TOGAF can help align business processes, data, applications, and IT infrastructure [43]. The architectural design developed in this study covers the provision of essential or basic

routine health services that always exist in the community, such as health examinations, disease treatment, and immunization [44]. The development of this PHR can also support national health priorities and the Germas program, which prioritizes promotive and preventive health efforts, especially for the prevention and control of noncommunicable diseases [45].

The application prototype in this study was developed as a mobile app. The increasingly widespread use of smartphone apps in the community and the ease of access to smartphones encourage the adoption of PHRs on mobile devices, or mobile PHRs [37]. In Indonesia, the number of smartphone users has reached 167 million, or 89% of the total population [46]. The results of the questionnaire in this study on the use of health applications section also showed that smartphones were the most popular devices for patients or individuals to access health applications in Indonesia. Smartphones also offer unique features such as a camera, GPS, and touch screen that can be used to extend the usefulness of mobile PHR, such as scanning and importing paper documents, recording certain symptoms, creating videos, or scanning bar codes for medical purposes [37].

Implications

As explained in the Introduction section, previous research on PHR design in low- and middle-income countries [6-9] has involved users in exploring the needs and usability of the PHR design. However, they did not explain the integration of the PHR with other health applications. Although there is a study describing PHR integration using the distributed PHR model, this study did not involve users or stakeholders in designing the model [10]. Our study complements the gaps in previous studies in low- and middle-income countries by designing an integrated PHR in Indonesia, which involves related stakeholders in requirement gathering and evaluation.

The theoretical implications of this research are the contribution to the field of PHR research by presenting design science as an approach for designing an integrated PHR system for a low- or middle-income country context that takes into account the specific characteristics of the Indonesian health care system. By using DSR, it can be ensured that the PHR model developed is based on scientific theory and methods. In addition, the DSR approach helps researchers understand existing health care systems and the needs of various stakeholders, such as patients, health care providers, and health regulators, in developing PHRs. DSR then includes evaluating the proposed PHR model, which can help identify any issues and make necessary improvements so that the designed PHR can follow the health system in Indonesia.

We developed architecture and application prototypes based on health systems in Indonesia, which comprise routine health services, including disease treatment and health examinations, as well as promotive and preventive health efforts. In addition, in this study, there are health referral functions that have not been discussed in the previous review of PHR functionality [5] and other studies in low- and middle-income countries [6-10]. This referral function is needed as the health system in Indonesia has a tiered referral program that should be followed by JKN patients [13].

The practical implication of this study is that this research is expected to be a guide for health regulators, health facilities, or health application vendors in designing an integrated PHR system in Indonesia. The architectural design in this study can provide an overview to integrate the PHR into the health services process in Indonesia, information about parties that need to be integrated into the PHR, and technology that can be used for PHR implementation. The prototype design in this study provides a guideline to implement PHR functions that focus not only on health care but also on disease prevention and health promotion.

Conclusions

The architecture design of the integrated PHR system in Indonesia refers to the TOGAF version 9.2, which is divided into 5 components: architecture vision, business architecture, application architecture, data architecture, and technology architecture. We developed a high-fidelity prototype for patients and physicians. The functionalities that were implemented were the priority functions defined in the architecture. The architecture evaluation stated that the architecture design had already described the needs and processes of health services in Indonesia as well as the technology needed for implementation. Improvements were made to the application architecture and data architecture to add the stakeholders that need to be integrated and the required functionality to the PHR. Prototype evaluation resulted in adding the necessary information to the functions that were developed, such as linking the medication order and messaging functions to the teleconsultation application and adding a health screening function. The limitation of this research is that the evaluation only focused on assessing the suitability of the integrated PHR model for the needs of health programs in Indonesia from the perspective of IT or eHealth experts. Future studies should be conducted to evaluate the prototype PHR from the perspective of patients and physicians as the primary users of the PHR.

Acknowledgments

This work was supported by a Pendidikan Magister Menuju Doktor untuk Sarjana Unggul grant from the Ministry of Research, Technology, and Higher Education, Republic of Indonesia, with research grant contract NKB-341/UN2.RST/HKP.05.00/2021.

Authors' Contributions

NCH designed the study, performed data collection, analyzed the data, and wrote the paper. PWH designed the study, provided writing assistance, and approved the final version to be submitted. ANH provided writing assistance and approved the final version to be submitted.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Interview guide.
[DOCX File , 20 KB - medinform_v11i1e44784_app1.docx]

Multimedia Appendix 2
COREQ (Consolidated Criteria for Reporting Qualitative Research) checklist.
[DOCX File , 26 KB - medinform_v11i1e44784_app2.docx]

Multimedia Appendix 3
Interview respondents.
[DOCX File , 23 KB - medinform_v11i1e44784_app3.docx]

Multimedia Appendix 4
Health application use.
[DOCX File , 25 KB - medinform_v11i1e44784_app4.docx]

Multimedia Appendix 5
Functionality codes.
[DOCX File , 23 KB - medinform_v11i1e44784_app5.docx]

Multimedia Appendix 6
Summary of the personal health record module and functionality based on the requirements of health organizations and patients.
[DOCX File , 24 KB - medinform_v11i1e44784_app6.docx]

Multimedia Appendix 7
Process flows in the personal health record.
[DOCX File , 231 KB - medinform_v11i1e44784_app7.docx]

Multimedia Appendix 8
Module and functionality in the personal health record.
[DOCX File , 25 KB - medinform_v11i1e44784_app8.docx]

Multimedia Appendix 9
Data categories with data groups and descriptions in the personal health record.
[DOCX File , 23 KB - medinform_v11i1e44784_app9.docx]

References

1. Demiris G. Consumer health informatics: past, present, and future of a rapidly evolving domain. Yearb Med Inform 2016 May 20;Suppl 1(Suppl 1):S42-S47 [FREE Full text] [doi: 10.15265/IYS-2016-s005] [Medline: 27199196]
2. Heart T, Ben-Assuli O, Shabtai I. A review of PHR, EMR and EHR integration: a more personalized healthcare and public health policy. Health Policy Technol 2017 Mar;6(1):20-25. [doi: 10.1016/j.hlpt.2016.08.002]
3. Alsahafi YA, Gay V. An overview of electronic personal health records. Health Policy Technol 2018 Dec;7(4):427-432. [doi: 10.1016/j.hlpt.2018.10.004]
4. Tang PC, Ash JS, Bates DW, Overhage JM, Sands DZ. Personal health records: definitions, benefits, and strategies for overcoming barriers to adoption. J Am Med Inform Assoc 2006;13(2):121-126 [FREE Full text] [doi: 10.1197/jamia.M2025] [Medline: 16357345]
5. Harahap NC, Handayani PW, Hidayanto AN. Functionalities and issues in the implementation of personal health records: systematic review. J Med Internet Res 2021 Jul 21;23(7):e26236 [FREE Full text] [doi: 10.2196/26236] [Medline: 34287210]

6. Minoletti S, Rapisarda R, Giraldo L, Grande M, Sommer J, Plazzotta F, et al. User-centered design of a pediatric vaccination module for patients. Stud Health Technol Inform 2019 Aug 21;264:1096-1100. [doi: 10.3233/SHTI190395] [Medline: 31438094]
7. Farinango CD, Benavides JS, Cerón JD, López DM, Álvarez RE. Human-centered design of a personal health record system for metabolic syndrome management based on the ISO 9241-210:2010 standard. J Multidiscip Healthc 2018 Jan 9;11:21-37 [FREE Full text] [doi: 10.2147/JMDH.S150976] [Medline: 29386903]
8. Farzandipour M, Nabovati E, Farrokhian A, Akbari H, Rezaei hasanvand F, Sharif R. Designing and usability assessing an electronic personal health record for patients with chronic heart failure in a developing country. Inform Med Unlocked 2021;27:100804. [doi: 10.1016/j.imu.2021.100804]
9. Kaboutari-Zadeh L, Azizi A, Ghorbani A, Azizi A. Designing and evaluating a mobile personal health record application for kidney transplant patients. Inform Med Unlocked 2022;30:100930. [doi: 10.1016/j.imu.2022.100930]
10. Abdulnabi M, Al-Haiqi A, Kiah ML, Zaidan AA, Zaidan BB, Hussain M. A distributed framework for health information exchange using smartphone technologies. J Biomed Inform 2017 May;69:230-250 [FREE Full text] [doi: 10.1016/j.jbi.2017.04.013] [Medline: 28433825]
11. Indonesia Overview. The World Bank. 2020. URL: https://www.worldbank.org/en/country/indonesia/overview [accessed 2020-04-10]
12. Population: South-Eastern Asia. Worldometer. 2020. URL: https://www.worldometers.info/population/asia/south-eastern-asia/ [accessed 2020-10-08]
13. Mahendradhata Y, Trisnantoro L, Listyadewi S, Soewondo P, Marthias T, Harimurti P, et al. The Republic of Indonesia Health System Review. World Health Organization. 2017. URL: https://apps.who.int/iris/bitstream/handle/10665/254716/9789290225164-eng.pdf?sequence=1&isAllowed=y [accessed 2021-11-25]
14. GERMAS - Gerakan Masyarakat Hidup Sehat. Kementerian Kesehatan Republik Indonesia. 2017. URL: https://promkes.kemkes.go.id/germas [accessed 2020-10-11]
15. Program Indonesia Sehat dengan Pendekatan Keluarga. Kementerian Kesehatan Republik Indonesia. 2017. URL: https://www.kemkes.go.id/article/view/17070700004/program-indonesia-sehat-dengan-pendekatan-keluarga.html [accessed 2022-10-11]
16. Farmalkes S. Promotif Preventif Kesehatan untuk Membentuk SDM Unggul menuju Indonesia Maju 2045. Direktorat Jenderal Kefarmasian dan Alat Kesehatan. 2020. URL: https://farmalkes.kemkes.go.id/2020/02/promotif-preventif-kesehatan-untuk-membentuk-sumber-daya-manusia-sdm-unggul-menuju-indonesia-maju-2045/ [accessed 2021-09-23]
17. Cetak Biru Strategi Transformasi Digital Kesehatan 2024. Kementerian Kesehatan Republik Indonesia. 2021. URL: https://dto.kemkes.go.id/Digital-Transformation-Strategy-2024.pdf [accessed 2021-12-20]
18. Hevner AR, March ST, Park J, Ram S. Design science in information systems research. MIS Q 2004 Mar;28(1):75-105. [doi: 10.2307/25148625]
19. Miah SJ, Hasan N, Hasan R, Gammack J. Healthcare support for underserved communities using a mobile social media platform. Inf Syst 2017 Jun;66:1-12. [doi: 10.1016/j.is.2017.01.001]
20. Peffers K, Tuunanen T, Rothenberger MA, Chatterjee S. A design science research methodology for information systems research. J Manag Inf Syst 2007;24(3):45-77. [doi: 10.2753/mis0742-1222240302]
21. Inan DI, Win KT, Juita R. mHealth medical record to contribute to noncommunicable diseases in Indonesia. Procedia Comput Sci 2019;161:1283-1291. [doi: 10.1016/j.procs.2019.11.243]
22. Miah SJ, Hasan N, Gammack J. Follow-up decision support tool for public healthcare: a design research perspective. Healthc Inform Res 2019 Oct;25(4):313-323 [FREE Full text] [doi: 10.4258/hir.2019.25.4.313] [Medline: 31777675]
23. Niemöller C, Metzger D, Berkemeier L, Zobel B, Thomas O, Thomas V. Designing mHealth applications for developing countries. In: Proceedings of the 24th European Conference on Information Systems. 2016 Presented at: ECIS '16; June 12-15, 2016; Istanbul, Turkey p. 1-14.
24. Alharbi I, Alyoubi B, Hoque MR, Almazmomi N. Big data based m-Health application to prevent health hazards: a design science framework. Telemed J E Health 2019 Apr;25(4):326-331. [doi: 10.1089/tmj.2018.0063] [Medline: 30192202]
25. Iyawa GE, Herselman M, Botha A. Building a digital health innovation ecosystem framework through design science research. In: Proceedings of the 2019 Conference on Next Generation Computing Applications. 2019 Presented at: NextComp '19; September 19-21, 2019; Mauritius p. 1-6. [doi: 10.1109/nextcomp.2019.8883650]
26. Mauka W, Mbotwa C, Moen K, Lichtwarck HO, Haaland I, Kazaura M, et al. Development of a mobile health application for HIV prevention among at-risk populations in urban settings in east Africa: a participatory design approach. JMIR Form Res 2021 Oct 07;5(10):e23204 [FREE Full text] [doi: 10.2196/23204] [Medline: 34617904]
27. Khanom N, Miah SJ. On-cloud motherhood clinic: a healthcare management solution for rural communities in developing countries. Pac Asia J Assoc Inf Syst 2020 Mar 22;12(1):60-85. [doi: 10.17705/1pais.12103]
28. Venable J, Pries-Heje J, Baskerville R. FEDS: a framework for evaluation in design science research. Eur J Inf Syst 2016;25(1):77-89. [doi: 10.1057/ejis.2014.36]
29. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007 Dec;19(6):349-357. [doi: 10.1093/intqhc/mzm042] [Medline: 17872937]

30. Bengtsson M. How to plan and perform a qualitative study using content analysis. NursingPlus Open 2016;2:8-14. [doi: 10.1016/j.npls.2016.01.001]
31. Sessions R, DeVadoss J. A Comparison of the Top Four Enterprise Architecture Approaches in 2014. Microsoft Corporation. 2014 Oct. URL: https://download.microsoft.com/download/6/1/C/61C0E37C-F252-4B33-9557-42B90BA3E472/EAComparisonV2-028.PDF [accessed 2022-06-15]
32. Purnawan DA, Surendro K. Building enterprise architecture for hospital information system. In: Proceedings of the 4th International Conference on Information and Communication Technology. 2016 Presented at: ICoICT '16; May 25-27, 2016; Bandung, Indonesia p. 1-6. [doi: 10.1109/icoict.2016.7571907]
33. Haghighathoseini A, Bobarshad H, Saghafi F, Rezaei MS, Bagherzadeh N. Hospital enterprise architecture framework (study of Iranian University Hospital Organization). Int J Med Inform 2018 Jun;114:88-100. [doi: 10.1016/j.ijmedinf.2018.03.009] [Medline: 29673609]
34. The TOGAF® Standard, Version 9.2. The Open Group. 2018. URL: https://pubs.opengroup.org/architecture/togaf9-doc/arch/ [accessed 2021-11-25]
35. Herdiana O. TOGAF ADM planning framework for enterprise architecture development based on health minimum services standards (HMSS) at Cimahi city health office. IOP Conf Ser Mater Sci Eng 2018 Sep 26;407(1):012167. [doi: 10.1088/1757-899X/407/1/012167]
36. Harahap NC, Handayani PW, Hidayanto AN. Barriers and facilitators of personal health record adoption in Indonesia: health facilities' perspectives. Int J Med Inform 2022 Mar 22;162:104750. [doi: 10.1016/j.ijmedinf.2022.104750] [Medline: 35339888]
37. Kharrazi H, Chisholm R, VanNasdale D, Thompson B. Mobile personal health records: an evaluation of features and functionality. Int J Med Inform 2012 Sep;81(9):579-593. [doi: 10.1016/j.ijmedinf.2012.04.007] [Medline: 22809779]
38. Kyselov VB. Software tools. In: Kyselov VB, Domnich VI, Medvediev MH, Muliava OM, editors. Software Production and Game Modeling Methods. Lviv, Ukraine: Liha-Pres; 2019:5-35.
39. McGilvray D, Thomas G. Data categories. In: McGilvray D, editor. Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information. Cambridge, MA, USA: Elsevier Academic Press; 2008:39-44.
40. Application architecture. Oracle. 2022. URL: https://docs.oracle.com/cd/B10501_01/server.920/a96524/c07dstpr.htm [accessed 2022-03-02]
41. Saripalle R, Runyan C, Russell M. Using HL7 FHIR to achieve interoperability in patient health record. J Biomed Inform 2019 Jun;94:103188 [FREE Full text] [doi: 10.1016/j.jbi.2019.103188] [Medline: 31063828]
42. Hoque MR, Rahman MS, Nipa NJ, Hasan MR. Mobile health interventions in developing countries: a systematic review. Health Informatics J 2020 Dec;26(4):2792-2810 [FREE Full text] [doi: 10.1177/1460458220937102] [Medline: 32691659]
43. Kurnia S, Khoir S, Fuad A, Sanjaya GY, Dilnutt R, Achmad L, et al. One Data: COVID-19, health data connectivity and integration in Indonesia-a case study of Yogyakarta. The Australia-Indonesia Centre. 2021 Aug. URL: https://tinyurl.com/2ctfxxyh [accessed 2022-01-31]
44. Rokom. Pelayanan Kesehatan Essensial tetap Menjadi Prioritas di Masa Pandemi COVID-19. Sehat Negeriku. 2021 Jan 19. URL: https://sehatnegeriku.kemkes.go.id/baca/rilis-media/20201007/2735324/pelayanan-kesehatan-essensial-tetap-menjadi-prioritas-masa-pandemi-covid-19/ [accessed 2022-06-20]
45. Peraturan Menteri Kesehatan tentang Rencana Strategis Kementerian Kesehatan Tahun 2020-2024. Kementerian Kesehatan Republik Indonesia. 2020. URL: https://peraturan.bpk.go.id/Home/Details/152564/permenkes-no-21-tahun-2020 [accessed 2022-06-20]
46. Hanum Z. Kemenkominfo: 89% Penduduk Indonesia Gunakan Smartphone. Media Indonesia. 2021 Mar 7. URL: https://mediaindonesia.com/humaniora/389057/kemenkominfo-89-penduduk-indonesia-gunakan-smartphone [accessed 2021-11-15]

Abbreviations

API: application programming interface
BPJS Kesehatan: Badan Pelaksana Jaminan Sosial Kesehatan or Social Security Agency for Health
COREQ: Consolidated Criteria for Reporting Qualitative Research
DSR: design science research
FHIR: Fast Healthcare Interoperability Resources
GH: General hospital
HR: health regulator
JKN: Jaminan Kesehatan Nasional or national health insurance
mHealth: mobile health
PH: private hospital
PHC: primary health care
PHR: personal health record
PMI: Palang Merah Indonesia or Indonesian Red Cross
Puskesmas: pusat kesehatan masyarakat or primary health centers

TOGAF: The Open Group Architecture Framework
VDR: vendor

Edited by C Lovis; submitted 12.12.22; peer-reviewed by C Asuzu, KM Kuo; comments to author 06.01.23; revised version received 19.01.23; accepted 20.01.23; published 14.03.23.

Please cite as:
Harahap NC, Handayani PW, Hidayanto AN
Integrated Personal Health Record in Indonesia: Design Science Research Study
JMIR Med Inform 2023;11:e44784
URL: https://medinform.jmir.org/2023/1/e44784
doi:10.2196/44784
PMID:36917168

Original Paper

Using a Clinical Data Warehouse to Calculate and Present Key Metrics for the Radiology Department: Implementation and Performance Evaluation

Leon Liman[1], MSc; Bernd May[2], Dr rer nat; Georg Fette[3], Dip (Informatics); Jonathan Krebs[1], MSc; Frank Puppe[1], Prof Dr

[1]Chair of Computer Science VI, Würzburg University, Würzburg, Germany

[2]Management und Beratung in der Medizin (MBM) Medical-Unternehmensberatung GmbH, Mainz, Germany

[3]Service Centre Medical Informatics, University Hospital of Würzburg, Würzburg, Germany

Corresponding Author:
Leon Liman, MSc
Chair of Computer Science VI
Würzburg University
Am Hubland
Würzburg, 97074
Germany
Phone: 49 9313189250
Email: leon.liman@uni-wuerzburg.de

Abstract

Background: Due to the importance of radiologic examinations, such as X-rays or computed tomography scans, for many clinical diagnoses, the optimal use of the radiology department is 1 of the primary goals of many hospitals.

Objective: This study aims to calculate the key metrics of this use by creating a radiology data warehouse solution, where data from radiology information systems (RISs) can be imported and then queried using a query language as well as a graphical user interface (GUI).

Methods: Using a simple configuration file, the developed system allowed for the processing of radiology data exported from any kind of RIS into a Microsoft Excel, comma-separated value (CSV), or JavaScript Object Notation (JSON) file. These data were then imported into a clinical data warehouse. Additional values based on the radiology data were calculated during this import process by implementing 1 of several provided interfaces. Afterward, the query language and GUI of the data warehouse were used to configure and calculate reports on these data. For the most common types of requested reports, a web interface was created to view their numbers as graphics.

Results: The tool was successfully tested with the data of 4 different German hospitals from 2018 to 2021, with a total of 1,436,111 examinations. The user feedback was good, since all their queries could be answered if the available data were sufficient. The initial processing of the radiology data for using them with the clinical data warehouse took (depending on the amount of data provided by each hospital) between 7 minutes and 1 hour 11 minutes. Calculating 3 reports of different complexities on the data of each hospital was possible in 1-3 seconds for reports with up to 200 individual calculations and in up to 1.5 minutes for reports with up to 8200 individual calculations.

Conclusions: A system was developed with the main advantage of being generic concerning the export of different RISs as well as concerning the configuration of queries for various reports. The queries could be configured easily using the GUI of the data warehouse, and their results could be exported into the standard formats Excel and CSV for further processing.

(JMIR Med Inform 2023;11:e41808) doi:10.2196/41808

KEYWORDS

data warehouse; electronic health records; radiology; statistics and numerical data; hospital data; eHealth; medical records

Introduction

Background

Examinations performed by the radiology departments of hospitals, such as creating X-ray, computed tomography (CT), magnetic resonance imaging (MRI), or ultrasound images, are fundamental for many kinds of clinical diagnoses. Therefore, optimizing the use of radiology is important for any clinician working with it as well as for any patient being examined there. Such optimization has several advantages for the hospital, such as shorter times patients need to stay there as well as the ability to perform more radiologic examinations. It also has advantages for the patient, such as shorter times to wait for the radiology appointment as well as reduced radiation exposure, if unnecessary repeated examinations of the same body region are avoided.

Objectives

This optimization requires a good overview of the various key metrics of radiologic services and their changes over time. A systematic approach for computing such metrics is building and using a radiology data warehouse. The main requirements for a radiology data warehouse solution are:

- Generic data import from the underlying radiology information system (RIS), for example, via an intermediate data format
- Tools for enriching the basic data with inferred data via a preprocessing step, which allows for more simple and compact queries on the data
- An expressive query language
- A comfortable graphical user interface (GUI) for the query language, including the ability to specify the resulting reports as tables, graphs, or a standard export format for further processing
- An efficient engine for answering queries and generating reports

These requirements are further explained in the following sections.

State of the Art

The relevance of calculating the key metrics of radiology data [1], as well as the types of metrics, that are most interesting for radiology exports [2,3] has already been described. In addition, the benefits of presenting such metrics in an easily understandable dashboard [4,5] have been explained. Although such solutions have been implemented for many different uses cases [6-9], all of them use a fixed interface to 1 or multiple specific hospital information systems and provide the user with only a fixed selection of predefined calculations. In other systems, the primary goal is to show data from individual patients [10-12], which only allows for a limited amount of filtering and no user-defined queries on the data. Other approaches use a data warehouse [13,14] to unite data from several (still fixed) hospital information systems into a unified representation and therefore allow for various user-defined queries to be executed but are missing a GUI for users to specify their queries and instead have their users either use Microsoft Excel or Structured Query Language (SQL; ISO/IEC Joint Technical Committee 1/Subcommittee 32/Working Group 3) for report creation. For importing data into a clinical data warehouse, more generic solutions exist [15,16] but without an option to calculate additional features during the import. This could make certain types of reports difficult or, depending on the query language of the clinical data warehouse, even impossible to create. These solutions are further discussed in comparison to the developed solution in the final section.

For hospitals whose data have been used during the development of this system, the state of the art for calculating key metrics of their radiology data was to do so manually in Excel. Although this allows for many different reports to be created, it has several drawbacks, which are also discussed in the final section. An intermediate result of this work has already been described [17]. This is described in more detail, together with the improvements in the final result, in the following sections.

Requirements

Generic Data Import Into a Data Warehouse

A radiology data warehouse primarily needs data of the examinations (type of modality, date and time of the request, execution, and documentation of each examination), relevant basic and radiologic patient data, the medical question for the examination as well as the radiologic diagnosis, and information about the radiologic equipment used. Since hospitals use many kinds of RIS, the use of an intermediate data format facilitates the data import and makes it independent of the internal data structure of the RIS. In this project, an Excel (or comma-separated value [CSV]) table was used as an intermediate format, in which the RIS data could be exported and from which it could then be imported into the data warehouse. If a hospital could only export its RIS data into JavaScript Object Notation (JSON) format (a proprietary one or a standard one such as Health Level Seven [HL7] Fast Healthcare Interoperability Resources [FHIR]), the relevant information from this format could also be converted into a table (using an Excel or a CSV file) that uses the structure described in the next section. All the hospitals whose data were used during the development of this system were only able to provide Excel exports of their RIS data.

Semantic Preprocessing of the Basic Data

To make queries on the radiology data easier, preprocessing of basic data is useful. Therefore, additional values were inferred from the basic data during the import into the data warehouse. Two types of preprocessing were necessary for this project: The first type was calculations performed by combining information from the basic data. Examples of such precomputed values are the difference between the time when an examination was requested and the time when it was performed as well as the time when the radiologic images were interpreted. This is usually not available in the RIS directly but can be easily computed from the individual time stamps. The second type was standardizations of the basic data. For example, the medical question for the examination could be either available as unstructured text using different wordings or as a hospital-specific code, which must be associated with a readable, standardized description, for example, by using a regular

expression during the import. As new kinds of queries are requested, additional data may be required. Because of this, another requirement is the ability to perform an incremental update of the data warehouse with just the new data instead of deleting and reimporting everything that has already been loaded into it.

Types of Queries and Query Language

The developed system should be able to support a wide range of different calculations. The calculations requested by the hospitals with whose data the system has been used so far could be separated into 5 different categories, which are described here:

- Patients, appointments, and examinations per modality: The most common metric was the number of patients, radiology appointments, and examinations in the radiology department for each modality. Additionally, these numbers were separated between inpatients and outpatients, the department of the hospital requesting an examination, the region of the body that was examined, or the shift during the day in which the examination took place. All these numbers were used to provide a general overview of the use of the radiology department.
- Use of radiologic devices: A radiology department usually has many different devices for different modalities as well as often multiple devices for a single modality. To better distribute examinations and clinicians on these devices, their use is 1 of the requested calculations. The metrics for this use included the number of examinations and patients per device. Furthermore, the time for each examination as well as the vacancy between examinations were evaluated.
- Length of a patient's stay in the hospital: Depending on the disease, different lengths of stay in the hospital are necessary. To evaluate whether patients were staying longer in the hospital than expected, which results in a reduced capacity for new patients, the actual stay times were compared with the ones suggested by clinical guidelines.
- Waiting times: Short waiting times are in the interest of both the patient and the clinician requesting a radiologic examination. Therefore, for each modality, the time between the request of an examination, the actual appointment in the radiology department, and the availability of the clinical findings after the examination were calculated and compared.
- Multiple examinations for the same question: To find the answer to a specific medical question, in many cases, 1 kind of radiologic examination works best. If such radiologic examination is performed directly by an experienced radiologist (who also verifies whether the requested examination makes sense for the medical question), the chances are high that only 1 examination is necessary to answer the medical question. However, if for 1 medical question, multiple examinations with the same or with different modalities are performed, the patient has increased radiation exposure and fewer radiology appointments are available for other patients. To measure this, first, all the different sequences of modalities for different kinds of medical questions were calculated. Afterward, the number of patients with such sequences were counted and compared. In addition, the total time for which a patient with such repeated examinations stays in the hospital was evaluated.

The query language used by the developed system must be able to support these kinds of queries as well as additional ones requested by the hospitals. This is also important for evaluating possible ways in which any of these metrics can be improved. For example, unnecessary multiple examinations can perhaps be explained by too few available devices for the modality recommended for a question or by missing staff to operate a device. To verify whether the measures taken are successful, the query language must also be able to analyze the change in the metrics over time.

A common set of queries for data saved in the same way furthermore allows for an easy comparison of the calculated number between different hospitals. In addition, as none of the mentioned categories depends on a specific hospital, all these calculations can be performed for any hospital (even in different countries) if it is able to provide the necessary data from its RIS.

A Comfortable GUI for the Query Language and the Result Specification

Although the query language should be usable in textual form, a GUI is also required to create queries in a graphical way and automatically create the corresponding textual queries. As with the query language, the GUI should also allow for the layout of the requested report to be specified. The results of queries should be shown to the user as a table or as a graph. Furthermore, the results should be exportable into the standard formats Excel and CSV so that they can be further processed.

Furthermore, the GUI should make the system (with a limited amount of training) usable by the clinicians themselves and therefore should not require any knowledge of computer science.

Efficiency Requirements

Importing data into the data warehouse as well as creating reports using the query language on these data both should happen in a reasonable amount of time. For the initial import, the tool should not need longer than a few hours, and for querying the data, most of the queries should return their results in about 1 second, while more complex queries should not run for more than a few minutes. These requirements are necessary so that a user can quickly start to use the system and, while using it, easily try different variations of a query without a long waiting time for each result.

Methods

Ethical Considerations

In this paper only retrospective, pseudonymized patient data for patients with age groups below and above 18 years with a few attributes only about their radiologic examinations were used (dates, modality, device, localization, radiologic query, boolean values for insurance [statutory or private], boolean values for the type of stay in the hospital [inpatient or outpatient]). De-pseudonymization of the data was not possible for the authors of the study. Therefore, no ethics approval was necessary.

Concept

Processing Radiology Data for Importing It Into a Data Warehouse

The data from the RIS of the hospitals were provided to the tool as an Excel, CSV, or JSON file, in which each row represents a single examination. Each column in this file is 1 attribute, and the names of these attributes are written in the first row of the file.

To map these columns to attributes in the structure of the data warehouse, a configuration file (using Excel or CSV as well) was used. This file contained 1 row for each attribute and specified the name, identifier, and data type to use (eg, numbers or texts) when importing them together with the concrete values into the data warehouse. The columns containing the required metadata (eg, identifiers) must be specified in this configuration file as well.

As mentioned in the previous section, some values for the requested reports must be calculated based on the exported RIS data. To do this, several options were offered. Additional columns were added to the RIS export performing the calculations. These were then imported into the data warehouse like any other column in the RIS data by specifying them in the configuration file. The configuration file also provided an option to replace textual values with other values, which could, for example, be used to replace an abbreviation in the RIS export with a longer form. For more advanced calculations, several programming interfaces were offered and could be implemented for any value requiring such a calculation.

All the values from the RIS, together with the calculated values, were then saved to the data warehouse, and an index was created on them for increased query performance.

Creating Reports

As soon as all the needed values were saved in the data warehouse, queries on these data were run to calculate and create the requested reports. For this purpose, a query language was used to define the structure of the report. This was done by first specifying attributes to be queried as well as constraints on the values of these attributes. In the next step, these attributes were combined with the logical operators “and”, “or,” and “not.” These single attributes or groups of attributes were then used to specify the rows and columns of the requested report. For every combination of attributes in each row and each column, a query was created, resulting in the cells of the report. If additional constraints on all cells were required, other attributes were used to specify filters. Finally, the query language specified what type of count (examinations, appointments, or patients) should be returned. All this was either specified in textual form or graphically using the GUI of the data warehouse. To create a report, the query for each cell was run on the data and, depending on the configuration, the number of examinations, appointments, or patients was returned.

By using a query language like this, it is easily possible to run modifications of a query, which is further simplified by the ability of the data warehouse to save a query and load it again later.

After a query was configured and executed, the interface of the data warehouse showed the results as a table and provided the option to export this table into the standard formats Excel and CSV. The results of some predefined queries were also shown as graphics.

Implementation

The Clinical Data Warehouse PaDaWaN

PaDaWaN (short for Patient Data Warehouse Navigator) [18] was used as the clinical data warehouse. Its core is a database containing all the used medical information and a separate index to increase the speed of queries on the data. To specify the queries, PaDaWaN uses its own query language as well as its own web interface. Furthermore, it provides the ability to export and save query results. All these parts are described in more detail in the following sections.

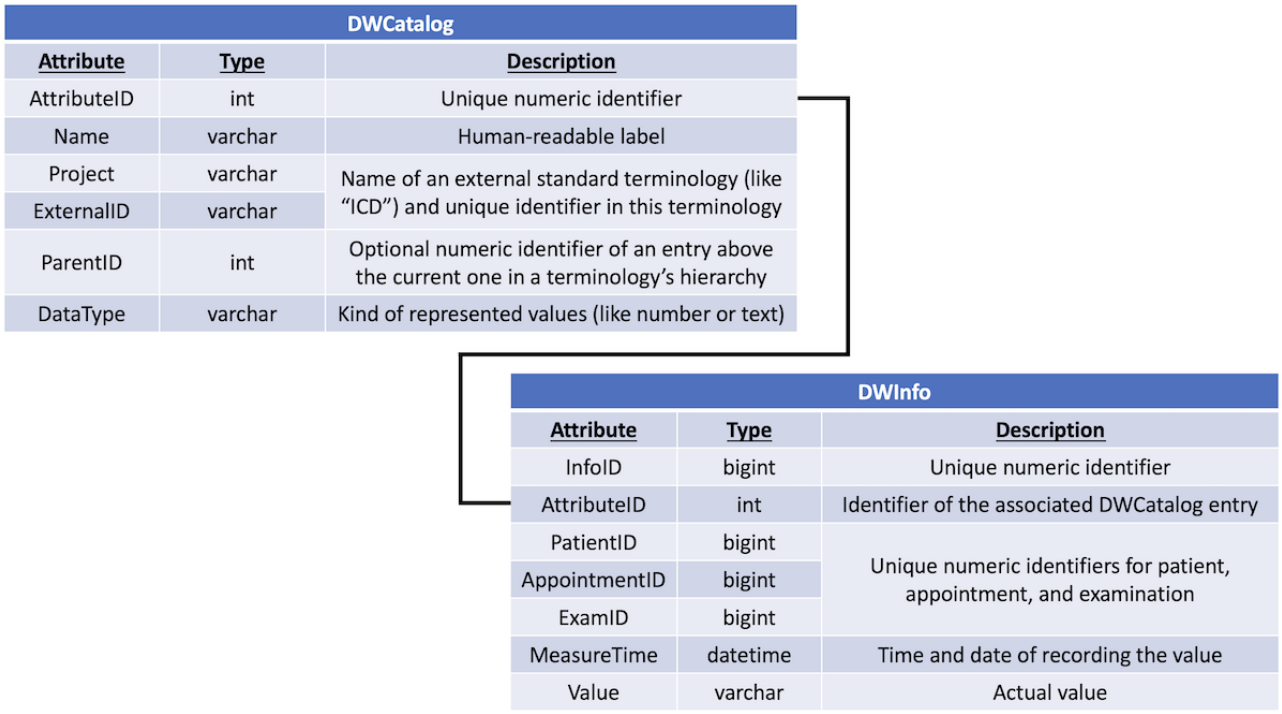
Database Structure

PaDaWaN stores its data in a database, which could be either a Microsoft SQL [19] or a MySQL [20] database. The table structure is based on the entity attribute value model [21]. It consists of 2 main tables, as shown in Figure 1.

The first table (DWCatalog) contains a catalog of all the possible types of information in PaDaWaN. This could represent, for example, different types of diagnoses or laboratory measurements. Each entry in this table is uniquely identified by a numeric AttributeID as well as by the combination of ExternalID (the ID in the terminology defining the entry, such as, the International Classification of Diseases [ICD] code [22] I50) and Project (the name of the whole terminology, such as “ICD”). AttributeID is automatically generated by the database and is only valid for a single installation of the system. As further explained later, only the combination of ExternalID and Project is used to identify entries from this table in a query, and therefore, only this combination must be unique among different systems if the same query should be used for all of them. For easier usage in the PaDaWaN interface, every entry has a readable name (eg, “heart failure”). The ICD terminology, for example, uses a hierarchal structure. To save this or any other hierarchy among the attributes, the ParentID field is used and contains the AttributeID of the entry, that is, the parent of the current entry. The kind of data (eg, Boolean, number, or text) that could be saved for an entry is specified with the DataType field.

In the DWInfo table, all the concrete patient data are saved. Each entry in this table is uniquely identified with an automatically generated numeric InfoID and is associated with a type of information from the DWCatalog table by its AttributeID. With the other ID fields, each entry in this table is also associated with a patient, appointment, and examination. The time and date on which a value has been recorded is saved in the MeasureTime field. The actual value (eg, the content of a patient’s discharge letter) is stored as text in the Value field.

Figure 1. Structure of the 2 main tables in PaDaWaN's database containing all the possible types of information (DWCatalog) and the information itself (DWInfo). ICD: International Classification of Diseases; PaDaWaN: Patient Data Warehouse Navigator.



DWCatalog

Attribute	Type	Description
AttributeID	int	Unique numeric identifier
Name	varchar	Human-readable label
Project	varchar	Name of an external standard terminology (like "ICD") and unique identifier in this terminology
ExternalID	varchar	
ParentID	int	Optional numeric identifier of an entry above the current one in a terminology's hierarchy
DataType	varchar	Kind of represented values (like number or text)

DWInfo

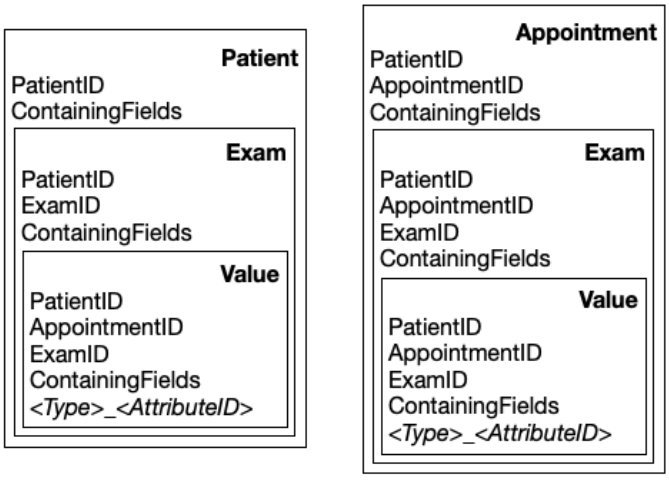
Attribute	Type	Description
InfoID	bigint	Unique numeric identifier
AttributeID	int	Identifier of the associated DWCatalog entry
PatientID	bigint	Unique numeric identifiers for patient, appointment, and examination
AppointmentID	bigint	
ExamID	bigint	
MeasureTime	datetime	Time and date of recording the value
Value	varchar	Actual value

Index Structure

To increase the speed of queries on the data in PaDaWaN's database, it was indexed with Apache Solr [23]. Solr saves data in documents, and the schema used for PaDaWaN is shown in Figure 2.

PaDaWaN offers the ability to search for data on 3 different levels: patients, appointments, and examinations. If a search is conducted on any of these levels, all patients/appointments/examinations should be found, containing all the requested combinations of attributes and values. To accomplish this in Solr, PaDaWaN uses Solr's feature to store documents nested in other documents. As shown in Figure 2, a document is created for each patient and each appointment. Another document is created for each examination and is stored inside the patient and appointment documents. Finally, for every value, another document is created and stored inside both examination documents. Although this approach requires more disk space as each value is saved twice, this greatly increases the speed of queries being run on the patient level compared to a document structure, where the appointments are nested inside the patient documents. Each document contains all the available IDs, as described in the previous section. Additionally, all of them contain a field named ContainingFields, which stores the AttributeIDs of all values contained in the current document. This allows a query to restrict the number of top-level documents it must search on for concrete values of the attributes. These values are stored in a field in the value document, whose name is generated by combining the attribute's type with its ID.

Figure 2. Document structure of PaDaWaN's Solr index using nested documents for examinations and values in separate parent documents for patients and appointments. In addition to the numeric identifiers for patients, appointments, and examinations under ContainingFields, all AttributeIDs of all values inside a document and its children are saved. The values themselves are stored in dynamic fields named with a combination of their type and AttributeID. PaDaWaN: Patient Data Warehouse Navigator.



Query Language

To specify the structure of the requested tabular result, PaDaWaN uses its own query language called Medical XML Query Language (MXQL). In the following example, the result table contains 2 rows and 2 columns. The rows contain 2 different types of modalities (X-ray and CT scan), and the columns contain 2 regions of the human body (abdomen and thoracic spine). For each combination of a row and a column, the number of matching patients from hospital A is returned. In the first cell, for example, the number of patients who undergo an X-ray examination of the abdomen is counted. This query is shown in MXQL in Figure 3 and in PaDaWaN's GUI in Figure 4. The result in Excel can be seen in Figure 5. This is a simple example used to explain the query language, PaDaWaN's GUI, and its export capabilities, and the results shown in Figure 5 may also be retrieved directly from an RIS (depending on its capabilities).

Each query in MXQL must contain at least the following 2 elements: Query and Attribute. Query is the root XML element and contains the whole rest of the query. Attribute contains information about the catalog entry whose values should be queried. To identify this catalog entry, Attribute uses the "domain" and "extID" properties, which map to the Project and ExternalID columns of the DWCatalog database table described before. The remaining elements of the query are optional and used for more complex queries. In the example query shown in Figure 3, the Attribute elements are further constrained to only match specific values for the catalog entries. This is done with the contentOperator and desiredContent properties, where desiredContent contains a value to be matched and contentOperator the way it should be matched. The IDFilter element is used to specify on which level all the attributes in the query should be combined. In the example in Figure 3, this is set to "PID," which means that all the attributes must have the same PatientID and that the number of matching patients should be returned by such a query. The last remaining elements of the query are DistributionRow, DistributionColumn, and DistributionFilter. They are used to return counts of multiple combinations of Attributes in a single query. Each DistributionRow becomes a row in the created result, and similarly, each DistributionColumn becomes a column. The DistributionFilter can be used to apply further constraints on all the combinations of rows and columns. Finally, the displayName property of the Attribute element can be used to provide a name for the created rows and columns. Not shown in the example is the ability of MXQL to combine multiple attributes with the logical combinations "and" and "or," which could even be nested inside another combination. MXQL also allows for the logical operator "not" to be added to any attribute.

Here, only the MXQL features used for this project are described. A complete documentation of this query language (in German) can be downloaded from PaDaWaN's website [24].

Figure 3. Sample query in PaDaWaN's query language MXQL. This query returns counts of patients (specified with the filterIDType "PID") for each combination of attributes specified as DistributionRows and DistributionColumns. In this example, the first combination would be all X-ray examinations of the abdomen. DistributionFilter restricts all the combinations to patients from hospital A. MXQL: Medical XML Query Language; PaDaWaN: Patient Data Warehouse Navigator.

```
<Query>
  <IDFilter filterIDType="PID">
    <DistributionRow>
      <Attribute domain="generated_attributes" extID="modality_group"
      displayName="X-Ray" contentOperator="EQUALS" desiredContent="X-Ray"/>
      <Attribute domain="generated_attributes" extID="modality_group" displayName="CT"
      contentOperator="EQUALS" desiredContent="CT"/>
    </DistributionRow>
    <DistributionColumn>
      <Attribute domain="generated_attributes" extID="body_region_group"
      displayName="Abdomen" contentOperator="EQUALS" desiredContent="abdomen"/>
      <Attribute domain="generated_attributes" extID="body_region_group"
      displayName="Thoracic Spine" contentOperator="EQUALS" desiredContent="TS"/>
    </DistributionColumn>
    <DistributionFilter>
      <Attribute domain="generated_attributes" extID="hospital_location"
      contentOperator="EQUALS" desiredContent="A"/>
    </DistributionFilter>
  </IDFilter>
</Query>
```

Figure 4. Web interface of PaDaWaN with the query from Figure 3. On the left side, the catalog of available attributes is shown and can be hierarchically expanded as well as searched. On the top of the right side, the query itself can be configured by dragging items from the catalog to create rows (Zeilen in German), columns (Spalten in German), and filters. With the 3 radio buttons in the top middle, the kind of IDs to be counted can be specified. Although meaning something else in German in this project, the buttons from left to right are used for patients, appointments, and examinations. The row of buttons in the middle are used to execute a query (Suchen in German) as well as to save and load queries (Speichern and Laden in German, respectively) and to export their results. After executing a query, the bottom right of the GUI shows its result (Ergebnis in German). The remaining buttons were not used for this project. The query shown here creates rows for X-ray (Rö: short form in German) and CT scan (CT: short form in German) examinations and columns for examinations of the abdomen and the thoracic spine (BWS: short form in German). The filter then restricts everything to just examinations from the hospital (Klinik in German) A. The remaining elements of the GUI were not used for this project. CT: computed tomography; GUI: graphical user interface; PaDaWaN: Patient Data Warehouse Navigator.

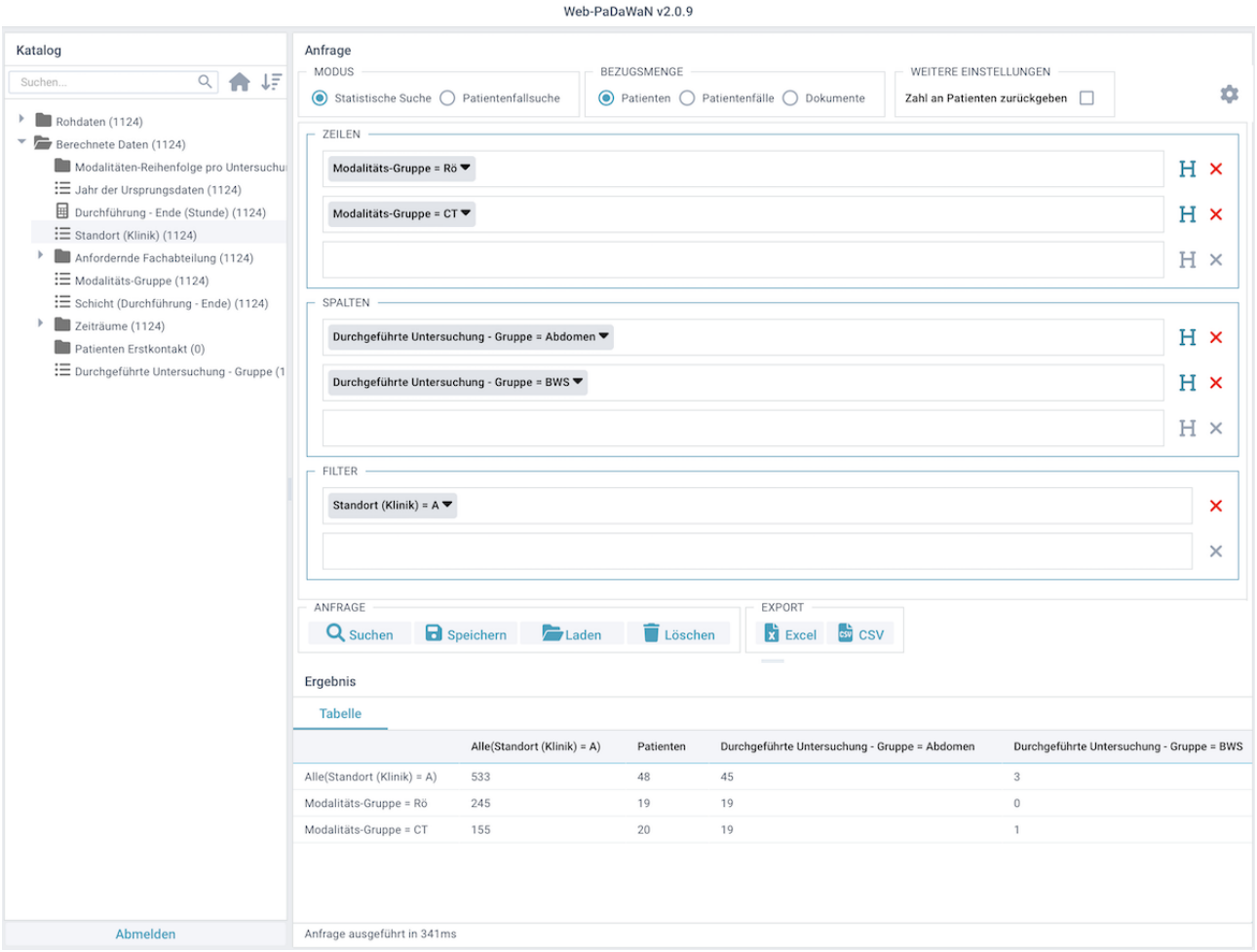


Figure 5. PaDaWaN Excel export for the query from Figures 3 and 4, where the number of patients with different kinds of radiology examinations (as rows) is counted for multiple regions of the human body (as columns). CT: computed tomography; PaDaWaN: Patient Data Warehouse Navigator.

	A	B	C
1		Abdomen	Thoracic Spine
2	X-ray	19	0
3	CT	19	1

Web Interface

PaDaWaN has its own graphical web interface allowing a user to search the available attributes, graphically configure a MXQL query, and preview the result table. The interface with the MXQL query from Figure 3 looks like Figure 4.

On the left side of the interface, the content of PaDaWaN's DWCatalog table (explained before) is shown and can be hierarchically expanded and searched. With the 3 radio buttons in the top middle of the GUI, the level on which all the attributes in the query should be found (for this project, patients, appointments, or examinations) can be configured. Via drag and drop, any attribute from the catalog can be placed in any section of the query to configure either rows (*Zeilen* in German), columns (*Spalten* in German), or filters (like that explained in the previous section). Finally, a configured query can be run by

pressing the Search button (*Suchen* in German). With the Save and Load buttons (*Speichern* and *Laden* in German, respectively), a configured query can be saved and any saved query can be loaded. The next 2 buttons provide the option to export a queried result in either Excel or CSV format. The bottom of the right half of the GUI shows the tabular result (*Ergebnis* in German) created after running a query. By clicking any of the attributes in the query on the right side of the GUI, a dialog box appears, where, for example, the value of an attribute that should be counted can be configured. In this example these values are *Rö* (short in German for X-ray), *CT*, *Abdomen*, *BWS* (short in German for thoracic spine), and *A* (for the name of the hospital; *Klinik* in German). All the remaining buttons were not used for the queries in this project.

Export of Query Results

PaDaWaN also offers an option to export query results in either Excel or CSV format using the Excel or CSV button, respectively, in Figure 4. When the query has finished, an Excel or a CSV file is created and offered as a download. When running the query shown in MXQL in Figure 3 and in the GUI in Figure 4, the Excel export looks like that in Figure 5.

As configured in the query, each row is a different kind of radiologic examination, and each column contains a different region of the human body. As the query was configured to return the number of matching patients, the first number means that in this (small and artificially generated) data set, 19 patients underwent an X-ray examination of the abdomen.

Export of the Radiology Data

For the developed system, data from a hospital's radiology department are needed. The system should be usable by many different hospitals with many kinds of RISs. Therefore, Excel, CSV, and JSON are used as the formats in which the RIS data can be exported and then imported from this file into PaDaWaN. As mentioned in the Introduction section, if a hospital is only able to provide its RIS data as a JSON file, this can also be transformed into a table and then saved as either an Excel or a CSV file. A part of an RIS Excel export is shown in Figure 6.

Figure 6. Sample of an RIS Excel export containing information about 1 examination in the radiology department per row and 1 attribute per column. CID: Case identifier; CT: computed tomography; RIS: radiology information system.

	A	B	C	D	E	F	G	H	I
1	Location	Device	Start of examination	CID	Examination-short	Examination-long	Insurance	Type of stay	Modality
2	Location 1	1-Xray	19-01-01 00:03	1	XCHPA	Chest radiograph I	private	outpat	Xray
3	Location 2	2-Xray2	19-01-01 00:21	2	XCHB	Bedside chest radi	private	inpat	Xray
4	Location 1	1-Xray	19-01-01 00:50	3	XWRL	Left wrist radiogra	statutory	outpat	Xray
5	Location 2	2-CT	19-01-01 01:21	4	CTSK	Skull computed to	private	outpat	CT
6	Location 2	2-Xray2	19-01-01 02:30	5	XCHB	Bedside chest radi	private	outpat	Xray
7	Location 2	2-Xray1	19-01-01 02:51	6	XHAL	Left hand radiogra	private	inpat	Xray
8	Location 1	1-Xray	19-01-01 02:52	7	XWRL	Left wrist radiogra	statutory	outpat	Xray
9	Location 1	1-Xray	19-01-01 03:20	8	XSHOL	Left shoulder radi	statutory	outpat	Xray
10	Location 1	1-Xray	19-01-01 03:20	8	XPEL	Pelvis radiography	statutory	outpat	Xray
11	Location 2	2-Xray1	19-01-01 04:40	9	XCHPA	Chest radiograph I	statutory	outpat	Xray
12	Location 2	2-Xray1	19-01-01 04:40	9	XABL	Abdomen radiogra	statutory	outpat	Xray
13	Location 2	2-Xray1	19-01-01 04:40	9	XABLL	Abdomen radiogra	statutory	outpat	Xray
14	Location 1	1-Xray	19-01-01 04:49	10	XT1R	Right big toe radio	private	outpat	Xray

Exporting uses a simple structure, where each row represents a single examination of a patient and each column contains 1 attribute with information about the examination. The title of the attribute must be given in the first row. The only required pieces of information are the ID of the patient's stay in the hospital, the start date and time of the examination, and the modality performed. All the remaining attributes could be different for each hospital, and the way they are imported into PaDaWaN is explained in the next section.

All the IDs that were used for this project had already been pseudonymized during the RIS export.

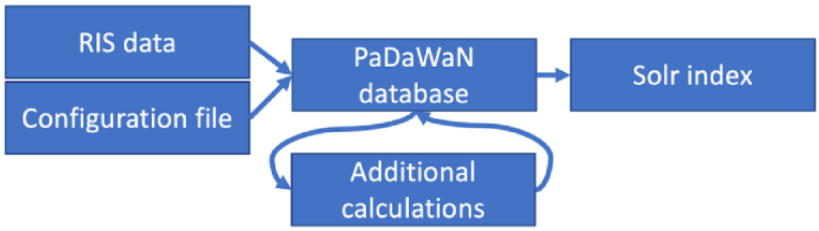
Import of the Radiology Data Into the Data Warehouse

The RIS export, as described in the previous section, was imported into PaDaWaN using the following steps:

- Step 1: A configuration file is created to specify the mapping of the RIS export columns to PaDaWaN catalog entries.
- Step 2: Using this configuration file, the data in the RIS export are converted to PaDaWaN database entries.
- Step 3: Additional precalculations are performed on the RIS data using an interface provided, and the results are saved in PaDaWaN's database as well.
- Step 4: A Solr index is created on these data.

An overview of this process is shown in Figure 7. Each step is described in more detail later.

Figure 7. Overview of the process for importing radiology data into the data warehouse. First, a configuration file is created and used to import the exported RIS data into PaDaWaN's database. On these data, additional precalculations are then performed. Finally, a Solr index is created for all the data in PaDaWaN's database. PaDaWaN: Patient Data Warehouse Navigator; RIS: radiology information system.



Configuration File to Map the Radiology Data to the Data Warehouse's Structure

An Excel (or CSV) configuration file was used to specify the mapping of the columns in the RIS export to the data structure of PaDaWaN. A configuration file for the data in Figure 6 would look like that in Figure 8.

The first row of this file must always look like that shown in Figure 8. Each of the following rows represents 1 column from the RIS export. If any of these columns should be ignored, they can be left out. The columns of the configuration file are used for specifying the PaDaWaN catalog entry that should be created (with the name from the DisplayTitle column and ExtID and DataType being directly used for database columns with the same names). The DataType "SingleChoice" is used for textual values with only a limited number of possible options (eg, modality). The ColumnName and ColumnNumber columns are used to identify a column in the RIS export. The ValueMappings column can be used to map abbreviations or codes in the RIS data to more readable names. Finally, the MetaDataType column is used to specify which columns contain which type of identifiers, the time an examination was performed, and the modality.

Figure 8. Sample of an Excel configuration file to specify the mapping between an RIS export and PaDaWaN's data structure. The ColumnName and ColumnNumber columns must match a column in the RIS export. The DisplayTitle, ExtID, and DataType columns are mapped to the corresponding columns in PaDaWaN's DWCatalog table. With ValueMappings, column abbreviations in the RIS export can be mapped to their longer form. The final column is used to specify which RIS column contains which type of metadata. CID: Case identifier; PaDaWaN: Patient Data Warehouse Navigator; RIS: radiology information system.

	A	B	C	D	E	F	G
1	**ColumnName**	**DisplayTitle**	**ExtID**	**DataType**	**ColumnNumber**	**ValueMappings**	**MetaDataType**
2	Location	Location	locatic	SingleCho	1		
3	Device	Device	statior	Text	2		
4	Start of exami	Start of exan	execSt	DateTime	3		MeasureTime
5	CID	Case identifi	cid	Text	4		CaseID
6	Examination-s	Code of exar	taskCc	Text	5		
7	Examination-l	Name of exa	taskTe	Text	6		
8	Insurance	Type of insu	payUn	SingleCho	7		
9	Type of stay	Type of stay	caseTy	SingleCho	8	inpat:inpatient;c	
10	Modality	Modality	modal	SingleCho	9		Modality

Import Process of the Radiology Data Using the Configuration File

When starting the import of the RIS export, first, the configuration file, as explained before, is read and then all the columns specified in the rows of the configuration file are imported into PaDaWaN.

For this, first, an entry in PaDaWaN's DWCatalog table is created with the values from the configuration file. To import concrete values from any column in the RIS export into PaDaWaN's DWInfo table, some metadata are required: PatientID, AppointmentID, and ExamID, as well as MeasureTime. These are specified with the MetaDataType column in the configuration file.

With the catalog entry and the metadata, each value in each column of the RIS export was saved into PaDaWaN's database.

Calculating and Importing Additional Values Based on the Radiology Data

As some calculations are not possible with PaDaWaN's query capabilities or would require complex queries, several interfaces (written in Kotlin [25]) are provided to specify additional calculations that should be executed during the RIS data import. Initially, for all these interfaces, the properties of the PaDaWaN catalog entry that should be created must be provided. Additionally, the RIS column names required for the calculation must be specified. The provided interfaces can then be used to either specify calculations that should occur for each examination (eg, calculating the shift during a day in which an examination was performed) or once for all examinations (eg,

to calculate sequences of examinations that have been performed for a single patient and for the same medical question with 1 or multiple modalities).

During the execution of all the implementations, the catalog entry specified by each implementation is created and the implemented methods to calculate the values and save them to the database are called.

Creating an Index on All Imported and Calculated Values

The last step during the import process of the RIS data is the creation of a Solr index on the data from PaDaWaN's database. For this purpose, all the values are fetched from the database and documents in the structure described before are created. These documents are then sent to Solr, which creates its index on them.

Incremental Updates of the Data Warehouse for New or Updated Radiology Data

The process of importing all the RIS data into PaDaWaN as well as creating a Solr index on it takes some time (shown in the next section). During the work on this project, additional calculations on the RIS data, updates on existing calculations, and additional information from the RIS were needed in many cases. The whole process described in the previous section could be run again, which resulted in most values being imported or calculated again, although they did not change.

Therefore, a separate configuration file could be given to the importer, containing just the names of the attributes from the RIS export or from the implemented interfaces, that had to be processed. When using this option, just the columns and calculations of these attributes are processed and saved to the database. Afterward, Solr's ability to perform atomic updates [26] is used. In this way, the whole documents do not have to be created and indexed again, but instead, only small parts for the updated or added attributes are deleted and then added with the new values to the existing documents.

Another possibility for new radiology data would be data from new patients. In this case, the additional data can be exported from the RIS into a separate file and then the whole import process described before can be run for just this file so that only the new data are added to the database and the index and no processing of the existing data must be done again. If a near-real-time evaluation of the data is requested by a hospital, this process can also be run immediately any time new data are added to the RIS.

Performing Calculations on the Data Using the Data Warehouse and Exporting Their Results

After the RIS data and any additional calculations on them are saved to and indexed by PaDaWaN, PaDaWaN's web interface is used to create and run queries on these data. The usage of the interface as well as the query capabilities have already been described in the section on PaDaWaN before.

The process of creating a new query involves first specifying all the attributes for rows and columns whose combinations should be counted in the result. Optionally, additional filters can be configured for all these combinations. Next, the user chooses whether the number of matching patients, appointments, or examinations should be returned. Finally, the query is run (the matching MXQL query is automatically created by the web interface), and its result is either shown directly in the GUI or is exported to an Excel or a CSV file.

For reusing queries, PaDaWaN also provides an option to save and load queries.

PaDaWaN's web interface uses a REST-based interface, which can also be used directly without the GUI. To do so, the query must be created as an MXQL string and can then be sent to the interface. When PaDaWaN has finished the execution of the query, the result can be received in JSON format or as an Excel or a CSV file.

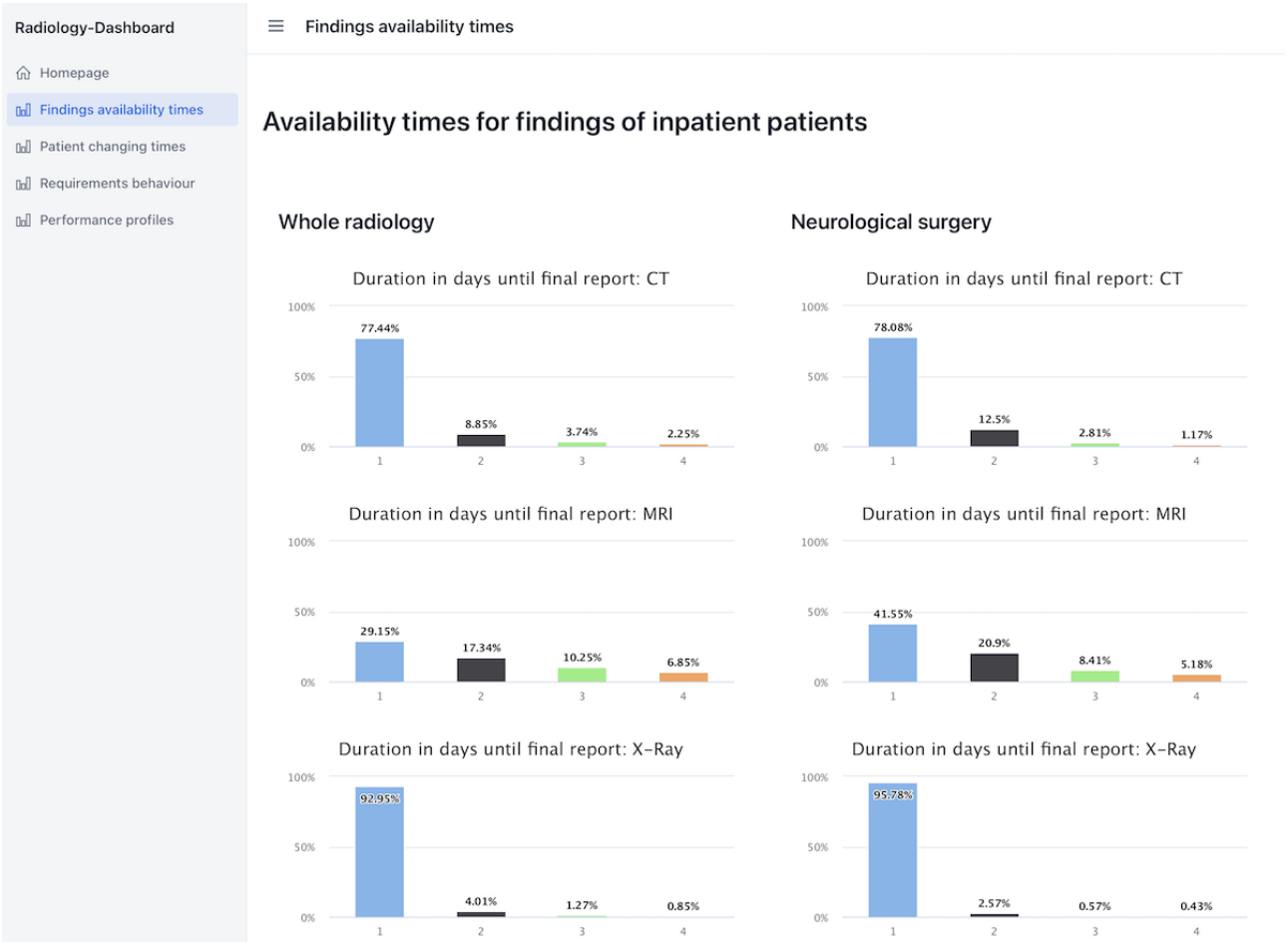
Presentation of Results Calculated by the Data Warehouse

As PaDaWaN allows for exporting of results into the standard formats Excel and CSV, these results can be easily imported by many different tools to perform further calculations or to create graphics. To present the 4 most common types of calculations for the hospitals involved in this project as graphics, a simple web dashboard was created and is shown in Figure 9.

The example calculates for all inpatients the percentage of findings for 3 modalities that has been available for 1, 2, 3, or 4 days after the examination in the radiology department. These numbers are further compared between the whole radiology department and just examinations that have been requested by the neurological surgery department.

To calculate these numbers, matching PaDaWaN queries were created and saved. When opening this dashboard, the queries were loaded and executed, and the numbers were extracted from PaDaWaN's result table.

Figure 9. Simple web dashboard with graphics for the 4 most common types of calculations on the radiology data of the hospitals involved in this project. The graphics in this figure show what percentage of findings for a radiological examination is available in up to 1, 2, 3, or 4 days. This is given for the 3 most common modalities and is compared between examinations in the whole radiology department and only the ones requested by neurological surgery. CT: computed tomography; MRI: magnetic resonance imaging.



Results

Technical Evaluation

In the following sections, details about the used data themselves as well as about the import and report creation process are presented.

Used Data

The developed system was tested with RIS exports from 4 different hospitals from different regions of Germany. Some of these data are shown in Table 1.

Table 1. Information about the used radiology exports of 4 different hospitals from Germany.

Details	Hospital A	Hospital B	Hospital C	Hospital D
Hospital sites, n	1	2	2	3
Time of data	2018	2019 to September 2020	2018 to June 2021	2019 to 2021
Patients, n	13,603	125,732	N/A[a]	N/A[a]
Appointments, n	28,886	384,186	307,174	241,148
Examinations, n	52,542	487,474	599,481	296,614
Values imported, n	2,50,001	11,650,688	15,014,221	5,339,024
Values generated, n	555,859	18,848,459	8,151,85	2,974,740
RIS[b] export size (MB)	14.9	75.7	56.1	21.7

[a]N/A: not applicable. The data from hospitals C and D contained no patient identifier, so the number of patients could not be specified.

[b]RIS: radiology information system.

The data were provided as an Excel export from the RISs of the hospitals. The last 3 hospitals had multiple radiologic sites in different cities. In addition, the time for which the data were exported was different, ranging from 1 year for the first hospital to 3.5 years for the third one. Only in the exports of the first 2 hospitals was a (pseudonymized) patient identifier included, so the number of patients could only be calculated for these 2 hospitals. Each value in a single cell of the RIS exports was imported into the data warehouse, and their number is specified in Table 1. For comparison, the number of values that were generated during the import is also specified. Numbers related to the import process are presented in the next section. Finally, in the last row, the size of the Excel files exported from the RIS is shown.

Process of Importing the Radiology Data Into the Data Warehouse

For the data of all 4 hospitals, a separate virtual machine (running in the internal network of the university of Würzburg) was created, and each of them was configured with 4 CPUs and 32 GB of RAM and stored on a solid-state drive (SSD). Inside of these machines was installed Ubuntu 20.04.4, together with MySQL 8.0.28, Java 11.0.14, and Solr 8.11.1. PaDaWaN's web interface was run on an Apache Tomcat [27] 10.0.18 server. On these virtual machines, the RIS exports were imported into PaDaWaN, resulting in the numbers shown in Table 2, which are discussed in the next section.

For 3 of the hospitals (the ones with data from multiple sites and years), the RIS export was provided as several Excel files, which were imported 1 by 1. Their number is specified in the first row of the first section of Table 2. After processing all files, the Solr index creation began.

The next row of Table 2 shows the total time needed for loading the RIS exports and saving their values to PaDaWaN's database. The number of attributes in the RIS export is specified in the last row of the first section.

Table 2. Numbers measured while importing radiology exports into the data warehouse.

Details	Hospital A	Hospital B	Hospital C	Hospital D
RIS data import				
Imported files, n	1	4	8	3
Import time (hours:minutes:seconds)	0:02:22	0:18:11	0:21:06	0:07:07
Imported attributes, n	45	26	11	18
Additional calculations				
Calculations, n	10	14	9	8
Time for each calculated attribute (seconds), mean (SD)	4.7 (2.9)	23.0 (30.2)	9.3 (2.3)	11.7 (8.0)
Total time (hours:minutes:seconds)	0:00:47	0:21:31	0:11:11	0:04:40
Index and database				
Solr index creation time (hours:minutes:seconds)	0:04:02	0:31:28	0:16:49	0:12:06
Time of the total process[a] (hours:minutes:seconds)	0:07:12	1:11:15	0:49:13	0:23:57
Database size (GB)	1.3	15.8	11.4	3.7
Solr index size (GB)	1.4	15.1	2.2	3.8
Incremental updates				
Examinations per day, mean (SD)	144 (61.8)	763 (288.5)	469 (200.3)	271 (222.3)
Time for adding these examinations[b] (hours:minutes:seconds)	0:00:18	0:01:29	0:01:00	0:00:28

[a]Total time for importing all attributes, calculating additional ones, and creating the Solr index.

[b]Time for incrementally adding just this average number of examinations per day.

In the second section of Table 2, numbers related to the additionally performed calculations are shown. These are the number of calculated attributes, the average time needed to calculate and save them to the database, and the total time for calculating and saving all these values.

The second-to-last section of Table 2 starts with the time needed to create a Solr index of all the imported values as well as the time needed for the whole import process of each hospital. In the last 2 rows, the size of the created database and Solr index is specified.

In the final section of Table 2, additional numbers related to the ability of the developed system to perform incremental updates are shown. Therefore, the average number of examinations per day for each hospital was calculated and then the time was measured to incrementally add just this number of examinations (along with additional calculations on them) to PaDaWaN's database and index.

Creating Reports on the Radiology Data With the Data Warehouse

Many reports were created using the data of all 4 hospitals, depending on the requirements of each hospital. Three reports

of different complexities, which were requested by most of the hospitals and were possible with the data provided by all of them, were created to show the time PaDaWaN needed to calculate those reports and export them as an Excel file. For each report, the number of matching examinations was restricted with MXQL to only include the data of 1 year. In all 3 reports, the 4 most common types of modalities for the hospitals (X-ray, CT, MRI, and ultrasound) were used. The following reports were created:

- Report 1: Number of examinations performed for the 4 modalities (as rows of the query) and for the types of stay in the hospital (inpatient or outpatient, as columns of the query)
- Report 2: Number of examinations performed for the 4 modalities (as rows of the query) and for the different hours of the day (from 8:00 a.m. to 7:00 p.m., as columns of the query)
- Report 3: Number of examinations requested by all the different organizational units of each hospital (as rows of the query) for the 4 modalities (as columns of the query)

The numbers related to the creation of these reports are shown in Table 3 and are discussed in the next chapter.

One Solr query was created for each possible row-column combination, which is why the number of executed Solr queries for each report equaled the number of rows multiplied by the number of columns. These numbers were identical for all 4 hospitals in the first 2 reports, as they used the same rows and columns. In the last report, 1 row was created for each organizational unit of the hospital, resulting in different numbers of rows for each hospital.

Table 3 also shows the average time in milliseconds Solr needed for each single query, as well as the total time to run all the Solr queries and export their results as an Excel file.

Table 3. Numbers related to the process of creating 3 reports of different complexities with the data warehouse.

Report and details	Hospital A	Hospital B	Hospital C	Hospital D
Report 1				
Rows, n	4	4	4	4
Columns, n	2	2	2	2
Solr queries, n	8	8	8	8
Time[a] (ms), mean (SD)	79.5 (111.6)	185.3 (264.0)	90.3 (117.3)	95.3 (180.7)
Total time	645 ms	1 s 489 ms	728 ms	767 ms
Report 2				
Rows, n	4	4	4	4
Columns, n	12	12	12	12
Solr queries, n	48	48	48	48
Time[a] (ms), mean (SD)	18.3 (7.7)	67.8 (23.3)	22.6 (12.4)	21.8 (13.7)
Total time	904 ms	3 s 275 ms	1 s 110 ms	1 s 69 ms
Report 3				
Rows, n	48	804	396	2054
Columns, n	4	4	4	4
Solr queries, n	192	3216	1584	8216
Time[a] (ms), mean (SD)	11.6 (4.2)	25.1 (3.2)	9.1 (3.4)	9.8 (3.3)
Total time	2 s 315 ms	82 s 233 ms	15 s 201 ms	84 s 758 ms

[a]Average time for the execution of each Solr query.

Comparison With the Creation of Reports Directly in Excel

Before using the developed tool, all 4 hospitals created such reports directly in Excel. To evaluate possible improvements compared to the report creation in Excel, this manual process was performed for new reports of different complexities with the largest data set (of hospital B) with the data of 1 year.

When the reports are created directly in Excel, nothing needs to be imported. Nevertheless, to simplify the calculations on the data, all the RIS exports of the considered year were combined into a single Excel file. The calculations otherwise executed during the import process were performed directly in Excel by using Excel formulas in new columns. As all these calculations were executed on each opening of the Excel file, all the RIS data, together with the calculated values, were then copied to another Excel file so that working with the data was faster.

The reports themselves could be created directly in Excel in many ways. If just single numbers are required, Excel's built-in filter capabilities can be used. To create the reports for this evaluation, Excel formulas were used to define the value of

each cell. These formulas were then copied to all the other cells, and their restrictions were adapted according to each row and column of the report that had to be created.

The results of this comparison are discussed in the next section.

User Feedback

Because the usage of the developed system consisted of various reports requested by the participating hospitals, their feedback was evaluated by describing the requests that could and that could not be created on the data provided by them.

In general, the requests could be divided into those of interest to all hospitals and special requests by an individual hospital. Of general interest was, for example, the number of patients, appointments, and examinations; the use of devices; repeated examinations of the same body region; and the waiting time for an examination. Specialized reports were mainly created for hospital B, which has the largest radiology department among the participating hospitals. The concrete reports that were created for each hospital are listed next.

Reports for Hospital A

- Specifically for patients with multiple myeloma or a hepatocellular carcinoma the number of patients for each modality and quarter of the year as well as for each type of stay in the hospital and each clinical department requesting a radiologic examination for such patients has been counted.
- For the same two types of diseases the number of patients with repeated examinations using the same or different modalities was counted.
- For two clinical departments requesting radiologic examinations the time between the request and the availability of the radiologic report has been evaluated.

Reports for Hospital B

Each of the following reports was requested for each site of the hospital as well as for regular radiology and neuroradiology:

- The number of patients, appointments, and examinations for each modality was counted. In separate reports, these numbers were further split by each shift and hour during the day or the body regions listed for the next report.
- For repeated examinations using the same or different modalities for the same body region, the number of patients was counted. The body regions of interest for this hospital were the abdomen, the cervical spine, the thoracic spine, the lumbar spine, the ankle joint, the knee joint, the hip joint, the shoulder joint, and the liver. In another report, the total time for these sequences of modalities was evaluated.
- For each modality, the time between the request for an examination and the actual start of it, as well as the availability of the radiologic report, was evaluated. In a similar way, the duration for just the examination itself was evaluated.
- The use of each radiologic device was evaluated by the duration for just the examination itself as well as the duration from the start of an examination until the start of the next examination using the same device.
- For another report, the number of patients at the radiology department for the first time or using each modality for the first time during a year was counted.
- The number of examinations was also counted for patients not older than 18 years and for the following special treatments: osteodensitometry, teleradiology, mammography, and nuclear medicine.
- Specifically for radiologic examinations of the spine, the difference between the actual time a patient stayed in the hospital in comparison to the time recommended by a clinical guideline was evaluated.

Reports for Hospital C

- The number of appointments and examinations for each modality was counted for each site of the hospital and each year. These numbers were calculated separately for each hospital department requesting a radiologic examination, for each shift during the day, for each type of stay in the hospital, and for each type of insurance a patient has.
- Only the number of MRI examinations was calculated for each hospital site and year separated by the following body regions: spine, abdomen, upper abdomen, pelvis, small intestine, joints, soft tissues of the neck, hand, foot, chest, skull, shoulder, and heart.

Reports for Hospital D

- For each site of the hospital, each year, type of stay in the hospital, and shift, the number of appointments and examinations was counted for each modality.
- Because for this hospital, the names of external private medical practices requesting radiologic examinations were also provided, the number of appointments and examinations requested by each of them was also counted for each hospital site, year, and modality.

With these reports, all the requests of hospitals A and B and some of the requests of hospitals C and D could be fulfilled. As mentioned before, hospitals C and D were not able to provide patient identifiers along with the rest of their exported data, and therefore, no number of patients and no sequences of multiple examinations (as they usually do not occur during the same appointment) could be calculated. The data provided by hospitals C and D also contained no time stamps except the time an examination started, and therefore, no time differences between, for example, the request for an examination and the actual appointment or the availability of the radiologic report, could be evaluated.

However, as long as the hospitals were able to provide the necessary data, all their requests could be fulfilled, and therefore, all of them were satisfied with the developed system.

Discussion

Principal Findings: Used Data

As shown by the numbers in the previous section, the imported data was diverse, with different numbers of years and attributes. The number of generated values was different as well (depending on the requested reports). However, some hospitals were not able to provide all the data for their requested reports,

such as hospitals titled C and D, which could not (or only with a lot of effort) provide a patient identifier, which resulted in the inability to create any reports with counts of patients. Nevertheless, due to its configurable and modular approach, the developed system can be used for these RIS exports as well, only requiring the creation of a new configuration file as well as some implementations of the interfaces for additional calculations. For hospital B, by far, the maximum number of reports was created, which resulted in the number of generated values exceeding the number of imported ones. One of these calculations, for example, was to count how many patients encountered multiple examinations for the same medical question with the same or different modalities. This directly pointed out multiple cases in which, for example, X-ray examinations had been conducted, followed by a CT or MRI examination, where only a CT or MRI examination would have been necessary, resulting in unnecessary radiation exposure for the patients as well as unnecessary radiology appointments.

Process of Importing the Radiology Data Into the Data Warehouse

When comparing the different numbers related to the import process, we found a correlation between the number of generated and calculated values and the time the developed tool needed to process them. However, even the RIS export with the maximum imported and calculated values (of hospital B) needed only about 1 hour 10 minutes for the whole process, making it fast to use even when installed in a new environment. For most reports, this time is only needed once, and multiple different reports can be created with the system afterward. If adaptions are needed (like for additional calculations during the import process), the mechanism for incremental updates can be used so that the time until the adaptions can be used for reports is even shorter. The storage required for the database and the Solr index together (31 GB for the largest data set of hospital B) can be easily found on many existing systems, and therefore, in most cases, no additional drives need be bought when using this tool. As shown in Table 2, the sizes of the database and Solr index were nearly identical for 3 of the hospitals. The difference between these sizes for hospital C was the reason that the data provided by it as well as the calculated values were mostly Boolean values. Although they are saved in similar form as other types of values in the database, the Solr index does not need to save and process any concrete textual or numeric value for them, resulting in the Solr index being a lot smaller than the database.

Creating Reports on the Radiology Data With the Data Warehouse

For many kinds of reports, the developed system can calculate and export them in a few seconds, as shown in Table 3. This allows a user to quickly iterate and try multiple configurations of a query. By using the preview option of PaDaWaN's interface, intermediate results do not always have to be exported to Excel or CSV. Even the third report in Table 3 could be created in less than 1.5 minutes for all 4 hospitals, although many single Solr queries were necessary for them. In all these reports, the queries were similar, resulting in the average time for each query becoming shorter with the total number of queries. Another observation from the created reports is that with a larger Solr index (hospital B has the largest one), the average time for each Solr query more than doubles compared to the reports created for the other hospitals but still goes down to 25 ms during the creation of the third (and largest) report.

Comparison With the Creation of Reports Directly in Excel

When comparing the developed tool with report creation directly in Excel, except for the combination of data from multiple Excel files into 1 (which is not necessary for the developed tool, because the RIS data are combined into 1 database and Solr index), no import of data are required, making this step faster and easier in Excel. However, for the calculation of additional values depending on the type of calculation, the required Excel formulas can get quite complex and are therefore more difficult to develop and maintain compared to a calculation written using the Kotlin method. To circumvent this disadvantage, Excel's ability to add scripts [28] can be used. During the creation of reports, the main disadvantage of using only Excel is the requirement for complex formulas, making the whole report more difficult to create and maintain compared to configuring a query in PaDaWaN's web interface. Especially the addition/deletion of an attribute to/from any row, column, or filter is easy in PaDaWaN's GUI, while this requires a user to adapt the Excel formulas in every cell. Therefore, PaDaWaN allows for easy ad hoc adaptation of reports even while discussing them with clinicians. Such ad hoc adaptation also benefits from the fast execution times of most PaDaWaN queries, as explained before. The drawbacks resulting from the use of Excel formulas for creating reports can be partially overcome by using Excel's feature to create pivot tables. This allows a user, in a similar way to PaDaWaN, to configure rows, columns, and filters of the requested table as well as the kind of numbers (eg, counts of patients or examinations) that should be returned. The disadvantages of this feature are that no logical combinations of attributes for a single column, row, or filter can be specified and would require additional precalculated columns. It also lacks PaDaWaN's ability for advanced searches on textual data directly as part of the query [18]. Altogether this evaluation showed that although most kinds of reports can be somehow created with Excel, especially more complex queries are difficult to configure and maintain there, while this can be done easily in PaDaWaN's GUI. A limitation of the developed tool is the requirement for an initial setup and for additional training of the clinicians on how to use it, while Excel is a tool that is already installed in many hospitals and many clinicians are already familiar with its usage.

Comparison With Alternative Solutions

In addition to the creation of reports directly in Excel, other solutions already exist that provide a user with key metrics on medical data. A comparison of the solutions introduced in the State of the Art section is presented in Table 4.

Table 4. Comparison of different existing solutions to calculate key metrics of radiology data (or of medical data in general in the last column).

Solution	Studies					
	[6-9]	[10]	[11]	[12]	[13,14]	[15,16]
Graphical results	Yes	No	Yes	Yes	No	Yes
Graphical query definition	No	Limited	Limited	Limited	No	Yes
User-defined queries	No	Limited	Limited	Limited	Yes	Yes
Additional calculations during import	N/A[a]	No	No	No	No	No
Import independent of a specific RIS[b]	No	No	Yes	No	No	Yes

[a]N/A: not applicable. Because these solutions operate directly on a radiology information system (RIS), no intermediate storage is used and therefore no additional attributes can be saved to it.

[b]RIS: radiology information system.

The first 4 solutions provide dashboards for the following uses cases: general imaging use [6], ordered and performed imaging studies for the emergency department [7], scheduled and in-progress examinations in pediatric radiology [8], and various metrics on orders, acquisition, interpretation, and reporting of radiologic images [9]. Although all of them could present their results as graphics, no additional queries (in addition to the ones predefined for the graphics) could be performed. Furthermore, the solutions only work with 1 or multiple specific RISs.

The next 3 solutions work in a different way, as their primary purpose is to display information about individual patients who are currently treated (or about to be treated) in the radiology department [10-12]. The values shown by them can be filtered, for example, by a specific type of examination, but no real queries on these data can be defined by the user. The graphics provided by Henkel et al [11] are limited to single patients and show, for example, the history of 1 of their laboratory values. Munbodh et al [12] also provide predefined graphics for the total number of examinations of different kinds during the past month. Although they all use intermediate storage for all the patients' data from the hospitals and RISs, no additional calculations on these data can be performed during the import, and of the 3, only the solution provided by Henkel et al [11] is not tied to a specific hospital or RIS.

The next 2 solutions use a data warehouse as a business intelligence tool for the radiology department [13] and combine radiology with pathology data [14]. Although both solutions allow for user-defined queries to be executed, these queries must be specified using Excel or SQL and not via a GUI. They also cannot create graphics from the query results and are tied to a specific RIS. Additionally, no precalculations on the RIS data can be performed and saved in the data warehouse.

The last 2 solutions are importers for i2b2 [15] and for i2b2 as well as PaDaWaN [16] and are not dependent on a specific RIS. They can use all the capabilities of these data warehouse solutions, including the ability for user-defined queries via a GUI and to show some predefined graphics based on these queries. However, because no additional calculations can be performed and saved to the data warehouse during the import, some kinds of reports are difficult or even impossible to create (eg, the evaluation of repeated examinations of the same body region). The developed system has the capability of additional calculations and therefore supports the most diverse kinds of reports.

Verification of the Calculated Metrics

As mentioned in the Introduction section, there are 2 main purposes of calculating all these metrics: From a patient's perspective, the waiting time for an appointment in the radiology department should be as short as possible and the exposure to radiation should be as low as possible. The hospital, however, wants to maximize its profits. By reducing the time until a diagnosis has been made, the patients can stay in the hospital for fewer days, which therefore allows the hospital to treat more patients. In addition, by eliminating or at least reducing unnecessary repeated examinations of the same region of the body, the radiology department also has the capacity to treat more patients. Treating more patients results in more profit for the hospital and shorter waiting times for patients. The reduction in unnecessary examinations also lowers the patients' exposure to radiation. Thus, both main purposes of the developed system can be achieved if metrics concerning the waiting times for appointments and diagnosis as well as repeated examinations can be calculated. The way this can be done has already been described before. For a hospital to be able to improve these metrics, it is also important that potential reasons for longer waiting times and repeated examinations be evaluated. Just from the RIS data themselves, this is, for example, possible by comparing the use of different radiologic devices to check whether the purchase of additional devices is necessary. Another possibility is the comparison of different hospital departments that request radiologic examinations. If a metric is significantly worse for one department compared to others, maybe the communication between this department and radiology needs to be improved. However, the developed system is not limited to RIS data alone. By combining these data with other data from the hospital information system in the data warehouse, additional possibilities for improvements can be found. For example, by comparing the diagnosis made by the radiology department with the final clinical diagnosis, the quality of the radiologic diagnosis can be evaluated. In addition, because the data warehouse can store data from multiple years, along with their time stamps, the developed system also supports the verification if any measure taken results in the desired improvement of specific metrics.

Limitations of and Bias in the Calculated Metrics

As all the metrics calculated by the developed system are based on data from the RIS and the hospital information system, their quality directly depends on the quality of the hospital's documentation. As this is not verifiable by the developed tool, it could only be assumed that this is done correctly. As mentioned in the previous paragraph, it is also important for the developed system that potential reasons for, say, longer waiting times for an appointment be evaluated. If, for example, a difference between the radiologic diagnosis and the final clinical diagnosis is not documented, it cannot be checked as a potential reason for longer waiting times or unnecessary examinations. In general, inferring conclusions from the calculated metrics can be difficult if potential causes are not documented. For example, if for a medical question, the ideal radiologic examination would be MRI, it may not be conducted, because it is too expensive or enough devices are not available. This could, for example, be further validated by comparing the performed examination with the one recommended by clinical guidelines (like the ones provided by the German Radiation Protection Commission [29]). Another potential bias in the calculated metrics may result from radiologic examinations that are performed externally (eg, at a private medical practice). If not properly documented, this is a missed indicator for the need for additional devices or employees. By combining RIS data with additional data from the hospital in the data warehouse, the text search capabilities of the data warehouse can potentially be used to find information about such external examinations in the patient's discharge letter. Additionally, the interpretation of the metrics could be different among different hospitals, resulting in limitations of the comparability of the metrics between them. Although, for example, a waiting time of a few days for an appointment in the radiology department could be acceptable for one hospital, it could be inacceptable for another one.

Conclusion

To summarize, the developed system has its main advantages in being generic concerning the export of different RISs as well as concerning the configuration of queries for various reports. To use it, the only requirement is the ability of an RIS to create an Excel, CSV, or JSON export. This can then be imported by creating a simple configuration file in Excel or as a CSV file as well. During the import process, additional values can be calculated by implementing several provided interfaces. If further values should be added later, this is easily possible with the ability to use incremental updates. Various reports with any combination of and restriction on the imported attributes can then be graphically configured using PaDaWaN's web interface. Finally, the results of these reports can be exported into the standard formats Excel and CSV so that they can be easily processed with many different tools.

The whole tool as a Docker [30] image, a sample RIS export, and a configuration file are publicly available on PaDaWaN's website [31].

The developed tool can in the future be further enhanced by, for example, adding the ability to calculate other numbers than the count of patients, appointments, or examinations, such as the average of numeric values found by a query. To improve the presentation of the results, the current ability to create graphics for some predefined reports can also be extended to be configurable by a user and therefore allow for the creation of many kinds of graphical reports.

Conflicts of Interest

None declared.

References

1. Georgiana V, Kartawiguna D. Evaluation of radiology data warehouse implementation on education, research, and quality assurance. 2016 Presented at: 2016 International Conference on Information Management and Technology (ICIMTech); November 16-18, 2016; Bandung, Indonesia p. 277-280. [doi: 10.1109/icimtech.2016.7930344]
2. Karami M, Safdari R. From information management to information visualization: development of radiology dashboards. Appl Clin Inform 2017 Dec 16;07(02):308-329. [doi: 10.4338/aci-2015-08-ra-0104]
3. Karami M. A design protocol to develop radiology dashboards. Acta Inform Med 2014 Oct;22(5):341-346 [FREE Full text] [Medline: 25568585]
4. Karami M, Safdari R, Rahimi A. Effective radiology dashboards: key research findings. Radiol Manag 2013;35(2):42-45. [Medline: 23638580]
5. Mansoori B, Novak RD, Sivit CJ, Ros PR. Utilization of dashboard technology in academic radiology departments: results of a national survey. J Am Coll Radiol 2013 Apr;10(4):283-288.e3. [doi: 10.1016/j.jacr.2012.09.030] [Medline: 23545086]
6. Halpern DJ, Clark-Randall A, Woodall J, Anderson J, Shah K. Reducing imaging utilization in primary care through implementation of a peer comparison dashboard. J Gen Intern Med 2021 Jan 03;36(1):108-113 [FREE Full text] [doi: 10.1007/s11606-020-06164-8] [Medline: 32885372]
7. Scheinfeld MH, Feltus W, DiMarco P, Rooney K, Goldman IA. The emergency radiology dashboard: facilitating workflow with realtime data. Curr Probl Diagn Radiol 2020 Jul;49(4):231-233. [doi: 10.1067/j.cpradiol.2020.02.013] [Medline: 32376121]
8. Shailam R, Botwin A, Stout M, Gee MS. Real-time electronic dashboard technology and its use to improve pediatric radiology workflow. Curr Probl Diagn Radiol 2018 Jan;47(1):3-5. [doi: 10.1067/j.cpradiol.2017.03.002] [Medline: 28533102]

9. Nagy PG, Warnock MJ, Daly M, Toland C, Meenan CD, Mezrich RS. Informatics in radiology: automated web-based graphical dashboard for radiology operational business intelligence. Radiographics 2009 Nov;29(7):1897-1906. [doi: 10.1148/rg.297095701] [Medline: 19734469]
10. Burns JL, Hasting D, Gichoya JW, McKibben B, Shea L, Frank M. Just in time radiology decision support using real-time data feeds. J Digit Imaging 2020 Feb 12;33(1):137-142 [FREE Full text] [doi: 10.1007/s10278-019-00268-2] [Medline: 31515754]
11. Henkel M, Horn T, Leboutte F, Trotsenko P, Dugas SG, Sutter SU, et al. Initial experience with AI Pathway Companion: evaluation of dashboard-enhanced clinical decision making in prostate cancer screening. PLoS One 2022 Jul 20;17(7):e0271183 [FREE Full text] [doi: 10.1371/journal.pone.0271183] [Medline: 35857753]
12. Munbodh R, Roth TM, Leonard KL, Court RC, Shukla U, Andrea S, et al. Real-time analysis and display of quantitative measures to track and improve clinical workflow. J Appl Clin Med Phys 2022 Sep 03;23(9):e13610 [FREE Full text] [doi: 10.1002/acm2.13610] [Medline: 35920135]
13. Prevedello LM, Andriole KP, Hanson R, Kelly P, Khorasani R. Business intelligence tools for radiology: creating a prototype model using open-source tools. J Digit Imaging 2010 Apr 15;23(2):133-141 [FREE Full text] [doi: 10.1007/s10278-008-9167-3] [Medline: 19011943]
14. Rubin DL, Desser TS. A data warehouse for integrating radiologic and pathologic data. J Am Coll Radiol 2008 Mar;5(3):210-217. [doi: 10.1016/j.jacr.2007.09.004] [Medline: 18312970]
15. Bauer C, Ganslandt T, Baum B, Christoph J, Engel I, Löbe M, et al. The Integrated Data Repository Toolkit (IDRT): accelerating translational research infrastructures. J Clin Bioinformatics 2015;5(Suppl 1):S6. [doi: 10.1186/2043-9113-5-s1-s6]
16. Fette G, Kaspar M, Dietrich G, Ertl M, Krebs J, Stoerk S, et al. A customizable importer for the clinical data warehouses PaDaWaN and I2B2. Stud Health Technol Inform 2017;243:90-94. [Medline: 28883177]
17. Liman L, Fette G, Krebs J. Calculating key figures for radiology departments using a clinical data warehouse ? A technical case report. Stud Health Technol Inform 2021;283:69-77. [doi: 10.3233/shti210543]
18. Dietrich G, Krebs J, Fette G, Ertl M, Kaspar M, Störk S, et al. Ad hoc information extraction for clinical data warehouses. Methods Inf Med 2018 May 25;57(S 01):e22-e29. [doi: 10.3414/me17-02-0010]
19. Introducing SQL Server 2022. Microsoft. URL: https://www.microsoft.com/en-gb/sql-server/ [accessed 2022-08-09]
20. MySQL HeatWave - one MySQL database service for OLTP, OLAP, and ML. MySQL. URL: https://www.mysql.com [accessed 2022-08-09]
21. Dinu V, Nadkarni P. Guidelines for the effective use of entity-attribute-value modeling for biomedical databases. Int J Med Inform 2007 Nov;76(11-12):769-779 [FREE Full text] [doi: 10.1016/j.ijmedinf.2006.09.023] [Medline: 17098467]
22. International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM). Centers for Disease Control and Prevention. URL: https://www.cdc.gov/nchs/icd/icd-10-cm.htm [accessed 2022-08-09]
23. Learn more about Solr. Apache Software Foundation. URL: https://solr.apache.org [accessed 2022-08-09]
24. Abfragesprache – Lehrstuhl für Künstliche Intelligenz und Wissenssysteme. Institut für Informatik - Universität Würzburg. URL: https://www.informatik.uni-wuerzburg.de/is/open-source-tools/padawan-data-query-tool/entwickler-sicht/abfragesprache/ [accessed 2022-08-09]
25. Kotlin v1.8.21. Kotlin Foundation. URL: https://kotlinlang.org [accessed 2022-08-09]
26. Apache Solr Reference Guide: updating parts of documents. Apache Software Foundation. URL: https://solr.apache.org/guide/8_11/updating-parts-of-documents.html#atomic-updates [accessed 2022-08-09]
27. Apache Tomcat. Apache Software Foundation. URL: https://tomcat.apache.org [accessed 2022-08-09]
28. Differences between Office Scripts and VBA macros. Microsoft. URL: https://docs.microsoft.com/en-gb/office/dev/scripts/resources/vba-differences [accessed 2022-08-09]
29. Recommendations for medical imaging procedures. Strahlenschutzkommission. 2020 Jun 22. URL: https://www.ssk.de/SharedDocs/Beratungsergebnisse_PDF/2019/2019-06-27Orientie_e.html [accessed 2023-03-27]
30. Develop faster. Run anywhere. Docker Inc. URL: https://www.docker.com [accessed 2022-08-09]
31. Download – Chair of Computer Science VI – artificial intelligence and applied computer science. Institut für Informatik - Universität Würzburg. URL: https://www.informatik.uni-wuerzburg.de/en/is/research/padawan-data-query-tool/download/ [accessed 2022-08-09]

Abbreviations

CT: computed tomography
GUI: graphical user interface
JSON: JavaScript Object Notation
MRI: magnetic resonance imaging
MXQL: Medical XML Query Language
PaDaWaN: Patient Data Warehouse Navigator
RIS: radiology information system

Edited by G Eysenbach; submitted 09.08.22; peer-reviewed by A R, Y Chu, M Alarifi; comments to author 19.12.22; revised version received 08.02.23; accepted 03.05.23; published 22.05.23.

Please cite as:
Liman L, May B, Fette G, Krebs J, Puppe F
Using a Clinical Data Warehouse to Calculate and Present Key Metrics for the Radiology Department: Implementation and Performance Evaluation
JMIR Med Inform 2023;11:e41808
URL: https://medinform.jmir.org/2023/1/e41808
doi:10.2196/41808
PMID:37213191

A Pragmatic Method to Integrate Data From Preexisting Cohort Studies Using the Clinical Data Interchange Standards Consortium (CDISC) Study Data Tabulation Model: Case Study

Keiichi Matsuzaki[1], MD, PhD; Megumi Kitayama[2], RN, MSc; Keiichi Yamamoto[3], PhD; Rei Aida[4], MSc; Takumi Imai[5], PhD; Mami Ishida[6], DPH, MD; Ritsuko Katafuchi[7,8], MD, PhD; Tetsuya Kawamura[9], MD, PhD; Takashi Yokoo[9], MD, PhD; Ichiei Narita[10], MD, PhD; Yusuke Suzuki[11], MD, PhD

1
2
3
4
5
6
7
8
9
10
11

Corresponding Author:
Keiichi Matsuzaki, MD, PhD

Abstract

Background: In recent years, many researchers have focused on the use of legacy data, such as pooled analyses that collect and reanalyze data from multiple studies. However, the methodology for the integration of preexisting databases whose data were collected for different purposes has not been established. Previously, we developed a tool to efficiently generate Study Data Tabulation Model (SDTM) data from hypothetical clinical trial data using the Clinical Data Interchange Standards Consortium (CDISC) SDTM.

Objective: This study aimed to design a practical model for integrating preexisting databases using the CDISC SDTM.

Methods: Data integration was performed in three phases: (1) the confirmation of the variables, (2) SDTM mapping, and (3) the generation of the SDTM data. In phase 1, the definitions of the variables in detail were confirmed, and the data sets were converted to a vertical structure. In phase 2, the items derived from the SDTM format were set as mapping items. Three types of metadata (domain name, variable name, and test code), based on the CDISC SDTM, were embedded in the Research Electronic Data Capture (REDCap) field annotation. In phase 3, the data dictionary, including the SDTM metadata, was outputted in the Operational Data Model (ODM) format. Finally, the mapped SDTM data were generated using REDCap2SDTM version 2.

Results: SDTM data were generated as a comma-separated values file for each of the 7 domains defined in the metadata. A total of 17 items were commonly mapped to 3 databases. Because the SDTM data were set in each database correctly, we were able to integrate 3 independently preexisting databases into 1 database in the CDISC SDTM format.

Conclusions: Our project suggests that the CDISC SDTM is useful for integrating multiple preexisting databases.

(JMIR Med Inform 2023;11:e46725) doi:10.2196/46725

KEYWORDS

data warehousing; data management; database integration; integrate multiple data sets; Study Data Tabulation Model; SDTM; Clinical Data Interchange Standards Consortium; CDISC

Introduction

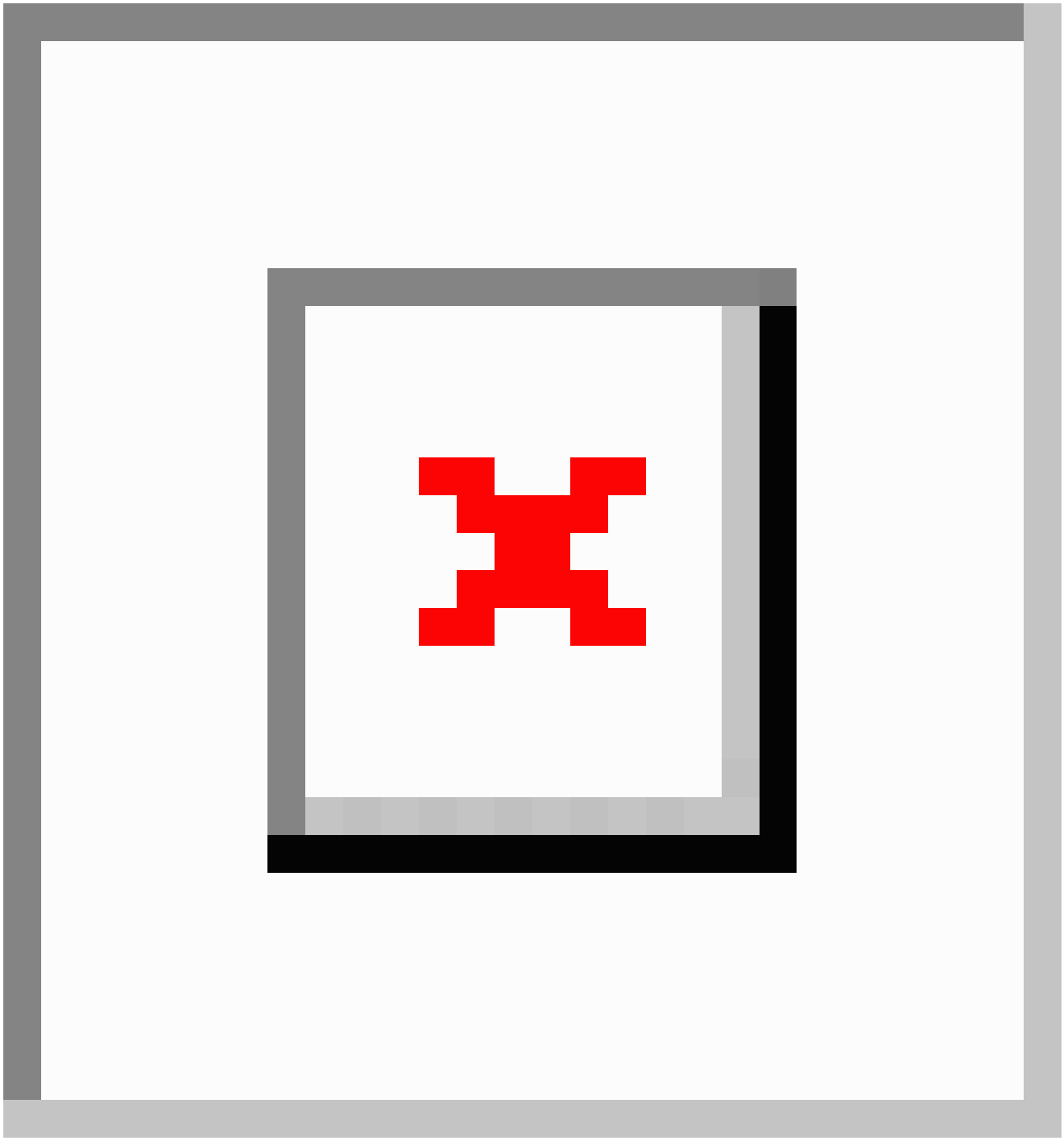
To use medical databases efficiently in clinical research, methods that efficiently integrate multiple databases must be established. The International Committee of Medical Journal Editors (ICMJE) requires researchers to include a data sharing statement when submitting a manuscript [1]. Moreover, there is a growing focus on the sharing of clinical research data and its uses. However, the current ICMJE statement makes no

mention of specific data standards for data sharing. Therefore, a discussion regarding specific ways to share data collected in clinical research is needed.

Recently, several medical societies and research groups have formed registries and conducted large cohort studies. The integration of databases with the same disease focus enables the analysis of data for many end points and patients. The reanalysis of data comprising large cohorts such as pooled analysis has statistical power and derives more reliable results [2]. For example, the Premenopausal Breast Cancer Collaboration, supported by the National Cancer Institute in the United States, published the results of several studies that used pooled analysis methods to integrate data from 20 independent cohort studies [3].

The Clinical Data Interchange Standards Consortium (CDISC) is a nonprofit, global organization that has developed several data standards to streamline clinical research [4]. The Study Data Tabulation Model (SDTM) is a data standard model for the sharing and integration of research data, which was initially developed to standardize the tabulation of clinical trial data submitted to the Food and Drug Administration (FDA) [5]. The concept of the CDISC SDTM is shown in Figure 1. The CDISC SDTM consists of several domains derived from clinical aspects, and each domain is identified by a unique 2-letter code [6]. Metadata are described in the data definition document named "Define" that is submitted with the data to regulatory authorities [7]. Each data item collected in different databases, using the SDTM and Define.xml, enables one to unify variable names and codes easily. Clinical research data warehouses using the CDISC SDTM are considered useful for data sharing in academic research.

Figure 1. Clinical Data Interchange Standards Consortium (CDISC) Study Data Tabulation Model (SDTM) concepts. eCRF: electronic case report form.



The integration of multiple data sets is difficult even among studies focused on the same disease. Major hurdles of data integration include the lack of standardization for the data format, variable names, and variable codes. Due to these problems, the manual conversion of data involves a large workload, which is likely to incur human error. Because the standardization of variable names and codes makes it easy to build a statistical data set, the CDISC SDTM provides a unique solution for database integration. However, many cohort studies have been conducted using a paper case report form (CRF) and formatted into data sets as a comma-separated values file or a spreadsheet file. It is difficult to convert these legacy data sets into the CDISC SDTM format because the variables need to refer to the CDISC variables and controlled terminology (CT).

Research Electronic Data Capture (REDCap) is an electronic data capture system developed by Vanderbilt University [8-10]. The “field annotation” function, introduced in REDCap version 6.5, can store meta-information for various standards related to clinical research, such as the CDISC, Systematized Nomenclature of Medicine (SNOMED), and Logical Observation Identifiers Names and Codes (LOINC). We previously developed “REDCap2SDTM,” a tool for parsing SDTM meta-information in the “field annotation” function and generating an XML file (Define-XML v2.0) with SDTM data [11,12]. This tool enables the efficient generation of SDTM

data from multiple preexisting research data sets, and it has been validated for SDTM data generation based on hypothetical clinical trial data. However, only a few data integration projects using an actual research data set were carried out [13-15].

The purpose of this project was to design a practical working model for integrating preexisting databases using the CDISC SDTM. Here, we report the pragmatic conversion of multiple preexisting databases based on the CDISC SDTM format.

Methods

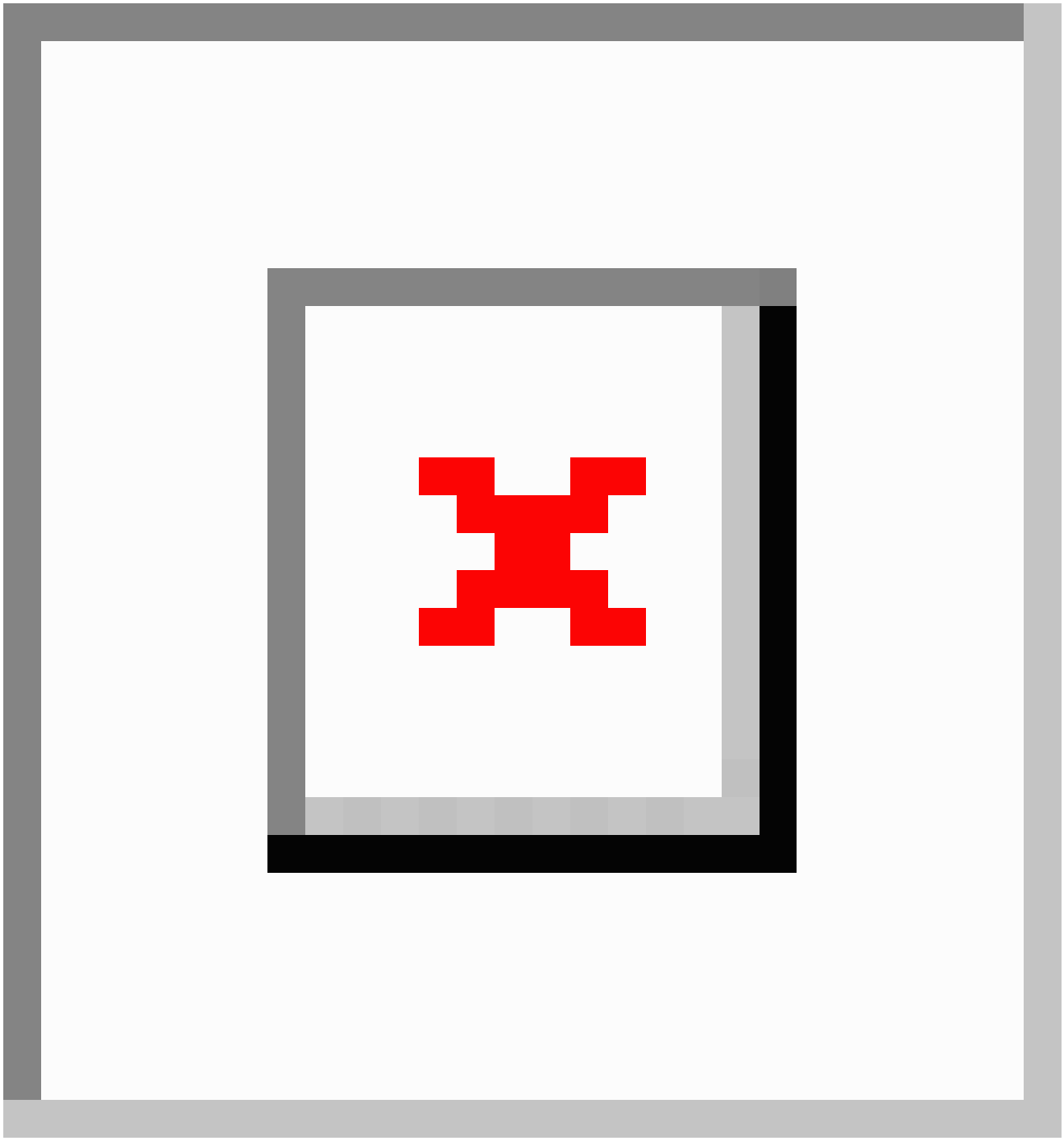
Ethical Considerations

Since this study was conducted on the structure of the database and not on patients, this study is outside the scope of ethical guidelines.

Project Structure

This project required multiple skill sets. A board-certified nephrologist (Japanese Society of Nephrology) with expertise in immunoglobulin A (IgA) nephropathy (including patient characteristics, laboratory data, and disease-specific items) confirmed the data structure in detail and constructed the independent database in REDCap. In parallel, a clinical data manager with CDISC SDTM expertise set the SDTM metadata in each variable. We outsourced the modification of REDCap2SDTM to a contract research organization to improve the efficiency of the SDTM data generation. The diagram of the study structure is shown in Figure 2.

Figure 2. Flowchart of the data integration. CT: controlled terminology; ODM: Operational Data Model; REDCap: Research Electronic Data Capture; SDTM: Study Data Tabulation Model.



Data Source

IgA nephropathy is the most common type of chronic glomerulonephritis in Japan. IgA nephropathy is a refractory disease in which 30% to 40% of patients reach end-stage renal failure after approximately 20 years [16]. Various clinical features and a chronic course are the hallmark of this disease; therefore, a database that collects multiple items and a prognosis is needed. To date, the IgA Nephropathy Working Group in Progressive Renal Diseases Research, affiliated with Research on Intractable Diseases from the Ministry of Health, Labor and Welfare of Japan, has conducted 3 cohort studies with over 1000 participants in each cohort. However, the collected items and data structures in each cohort study were not standardized, making the construction of an integrated database difficult. The number of collected items and the data structure of each cohort are shown in Table 1.

Table . Characteristics of each cohort studies.

Cohort	Items in the data set, n	Sites, n	Data structure
A	57	6	Vertical format: with repeated measurement data
B	65	6	Horizontal format: no repeated measurement data
C	582	42	Horizontal format: no repeated measurement data

Outline of Multiple Database Integration Work

The integration of multiple preexisting databases comprised the following three phases: (1) the confirmation of the variables in detail, (2) SDTM mapping, and (3) SDTM data generation and integration. The details of each phase are given below.

Phase 1: Confirmation of the Variables in Detail

In most cases, variable names differ by study, and the types of data also vary (date, digits, categorical variable, etc). Therefore, we set common values between each database in this phase.

Preexisting research data are stored in various formats between studies, including spreadsheets with a horizontal data (denormalized) structure. Since many SDTM domains are defined by a vertical data (normalized) structure, the data structure was transformed.

The main tasks of this phase were as follows:

- Standardize the variables in detail: code categorical data and nominal variables, unify date types (eg, YYYY/MM/DD), and improve the data format and the number of digits in clinical laboratory data in each data set
- Manage the data structure: transform repeated data from a horizontal structure to a vertical structure
- Validate the definition of variables: clarify data definitions and create a definition document in cooperation with specialists

Phase 2: SDTM Mapping

The CDISC has CT [17], and the terms used for each variable are specified in the *SDTM Implementation Guide* [6]. Through the use of CT, variables that were arbitrarily coded in different data sets can be derived as the same code. For example, if 1 data set coded male individuals as 1 and female individuals as 2 and another data set coded male individuals as 0 and female individuals as 1, the CT would code male individuals as “M” and female individuals as “F.” Therefore, the SDTM format data sets derived “M” for male individuals and “F” for female individuals. However, not all codes have specified CT, and coding lists for variables that are not specified must be created.

The domain model of the SDTM has a fixed domain of evaluation items to be stored. Therefore, each item in the data set must be mapped to the appropriate domain. For example, the “DM” domain contains the background of the patients (demographics), which includes age, sex, and race. The variable names were specified in each domain of the SDTM, for example, “SEX” for sex and “LBORRES” for laboratory results. Items with a unique code, such as sex, do not require a test code; the metadata are defined by the domain name “DM,” and the variable is named “SEX.” For items with various kinds of values, such as serum creatinine, a test code needs to be specified. For example, the meta-information of the creatinine test value must be defined by the domain name “LB,” the variable name “LBORRES,” and the test code “CREAT.” In addition, disease-specific end points are not defined in the standard domain of the SDTM. The SDTM does not allow new variables to be added arbitrarily; therefore, new variables must be defined in conjunction with the parent record using “supplemental qualifiers.” We determined the SDTM test code based on the appropriate code list from the SDTM CT.

Generally, in clinical studies, nominal scales (eg, male and female) are replaced by codes in the analysis. The method of assigning the code differs depending on the research and the data set, and recoding is necessary during database integration. We set both the domain and the meta-information of each data set based on the definitions confirmed in phase 1.

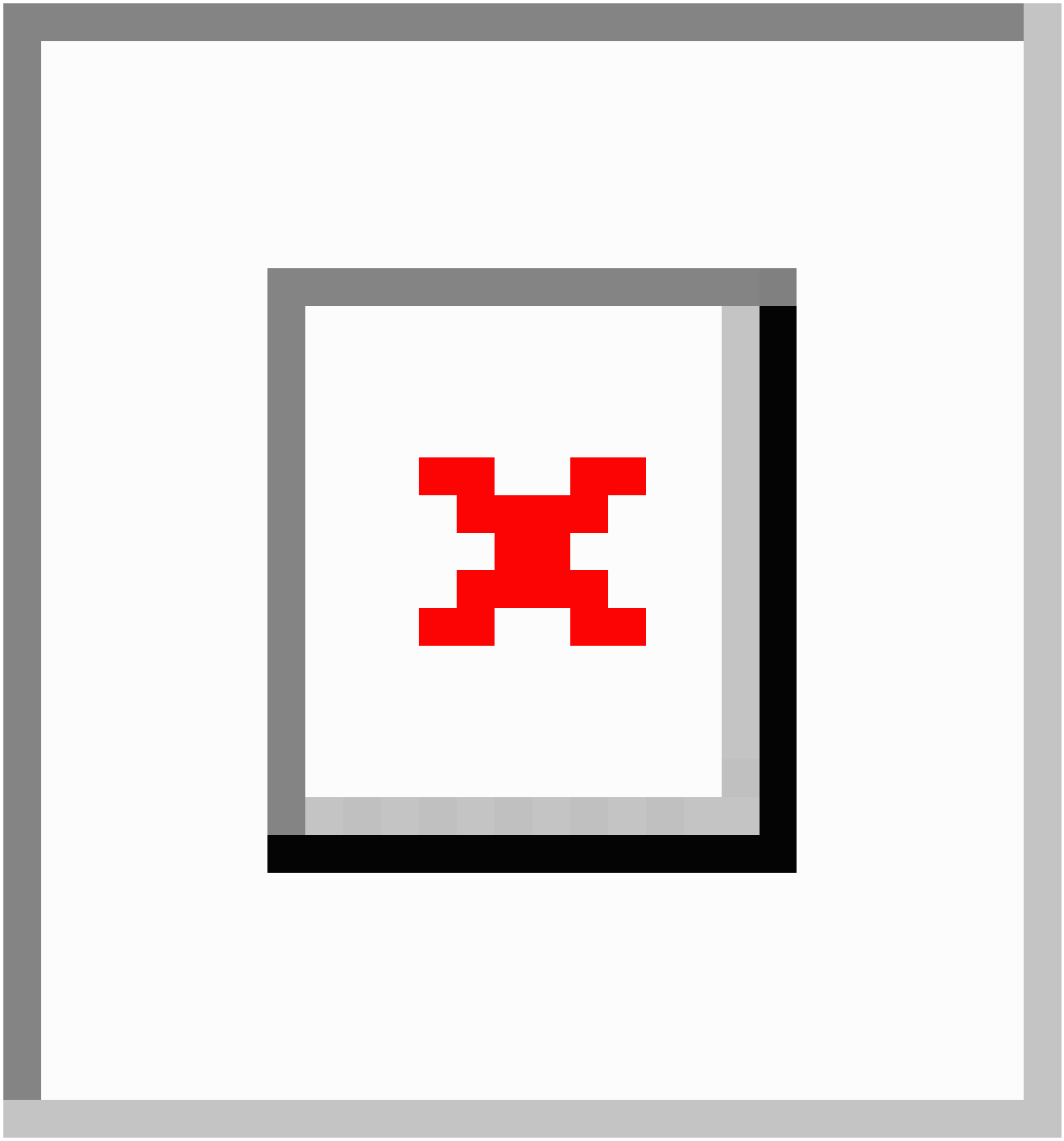
The main tasks of this phase were as follows:

- Recoding: map nominal variables and codes according to the CT or custom coding lists
- SDTM metadata mapping: map existing data variables to the SDTM domains

Phase 3: Generate SDTM Data in the Operational Data Model Format

In this phase, the SDTM metadata were manually set in the “field annotation” function (Figure 3). Subsequently, the data, including the data dictionary with the SDTM metadata, were downloaded in the Operational Data Model (ODM) format with SDTM metadata. Finally, REDCap2SDTM automatically generated each data set in the ODM format with the SDTM metadata.

Figure 3. Screenshot of “field annotation.” SDTM: Study Data Tabulation Model.



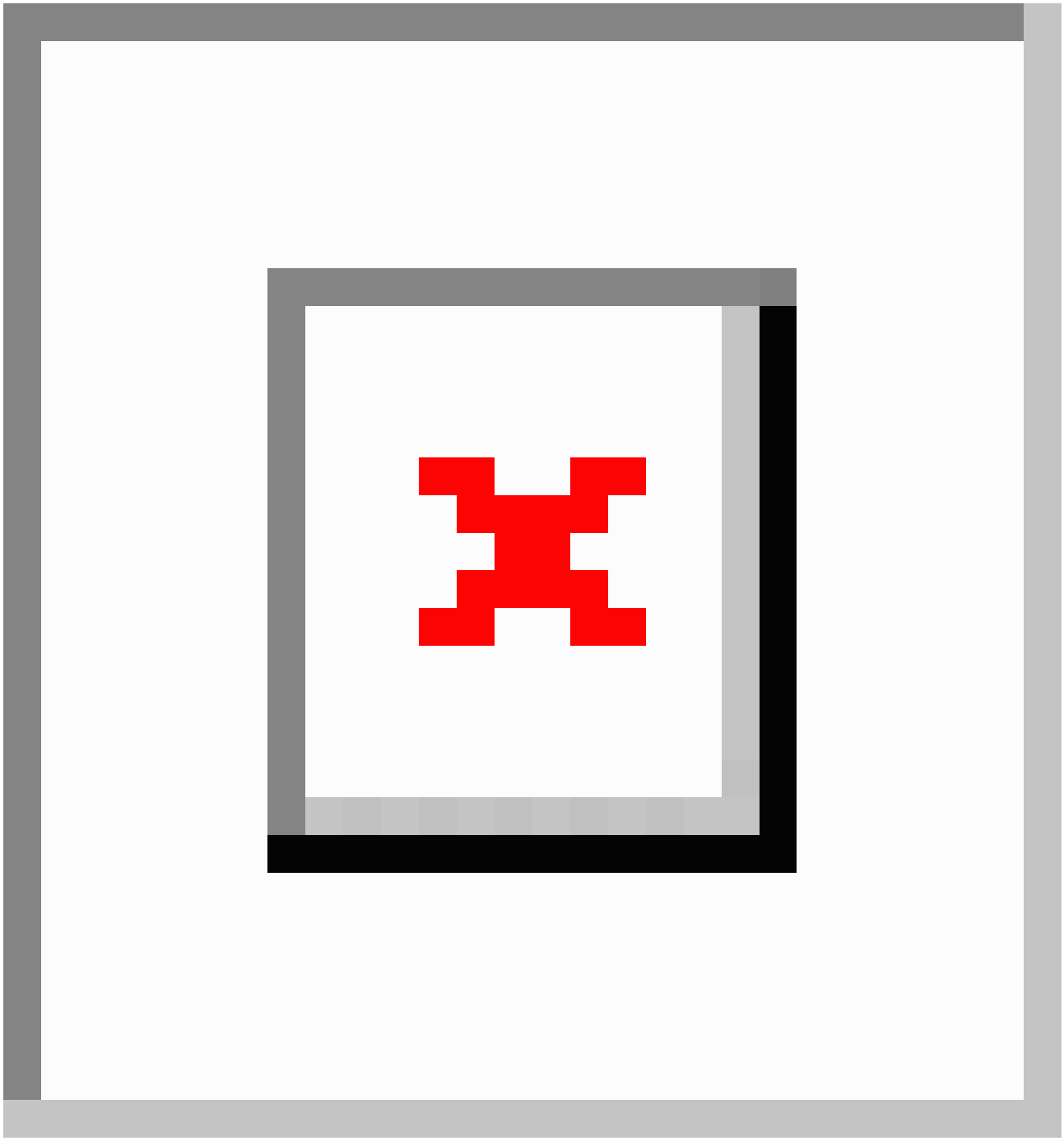
Common items in all data sets must be assigned the same metadata. Therefore, it is necessary to identify common items in all data sets to confirm the consistency of the metadata.

The main tasks of this phase were listed as follows:

- Check the consistency of the metadata: unify common items between each data set
- Generate SDTM metadata from each database: download and record data and data dictionaries and generate SDTM metadata with REDCap2SDTM
- Output SDTM metadata in the ODM format: retrieve the SDTM metadata output from REDCap2SDTM

A summary of the data integration process is shown in Figure 4.

Figure 4. Diagram of the study structure. CDISC: Clinical Data Interchange Standards Consortium; ODM: Operational Data Model; REDCap: Research Electronic Data Capture; SDTM: Study Data Tabulation Model.



REDCap2SDTM Version 2

REDCap2SDTM combined formatted ODM data that were embedded with 3 pieces of metadata (ie, SDTM domain name, variable name, and test code) into the field annotation of REDCap as the metadata corresponding to the variable name of the data set, to convert the database into the SDTM format. This tool dynamically generates SDTM data and a Define.xml file by parsing. The syntax of the meta-information is the CDISC Define-XML version 2.0 “ItemDef element.” REDCap2SDTM version 2 parses the object identifier attribute value and uses that information for mapping (eg, “IT.VS.VSORRES. SYSBP” and “IT.AE. AETERM”) [11,12].

The CDISC ODM is a vendor-neutral, platform-independent data format for exchanging and storing clinical research data and metadata that can be shared between different software systems [18]. In this case, we modified REDCap2SDTM to adopt the CDISC ODM format (REDCap2SDTM version 2; Multimedia Appendix 1). Due to this modification, REDCap2SDTM version 2 could convert the SDTM data to the ODM format and could expand to handle variables across multiple domains.

Results

This project was conducted from July 2018 to January 2019. Items were selected for integration in the SDTM metadata based on the opinions of the board-certified nephrologist.

Regarding disease-specific items, the histological classification of the disease was defined in the "SUPPMH" domain; items related to family history were defined in the "SUPPDM" domain, and the number of steroid pulse therapies was defined in the "SUPPCM" domain. The following domains were generated in this study: "DM" (demographics), "CM" (concomitant medications), "LB" (laboratory test results), "VS" (vital signs), "SUPPCM," "SUPPDM," and "SUPPMH."

The preexisting database included 57 total items for cohort A, 65 total items for cohort B, and 582 total items for cohort C. The metadata were set for 40 items for cohort A, 18 items for cohort B, and 102 items for cohort C. We found 17 common items. Finally, a total of 119 items were set for the SDTM metadata. Of these, 56 items used the nominal scale, 48 items could be recoded using CT, and 8 items required independently created code lists. Disease-specific items, such as the pathological classification based on the clinical guidelines for IgA nephropathy in Japan [19] and the Oxford classification [20], required their own code list. Table 2 lists the SDTM metadata of key items.

The data dictionary and ODM data were outputted from REDCap, and REDCap2SDTM version 2 was used to output the data in the SDTM format. The items defined by individual names in each database were collated based on the metadata by the CDISC SDTM.

Table . SDTM[a] metadata of the key items.

Items		Cohort A	Cohort B	Cohort C	SDTM metadata
Sex		Sex	Sex	Sex	DM.SEX
Birthday		Birth_date(Y/M/D)	Birth_date(Y/M/D)	birth date	DM.BRTHDTC
Age		__[b]	—	Age	DM.AGE
Vital sign					
	Systolic blood pressure	sbp_bx	sbp_bx	Sbp	VS.VSORRES.SYSBP
Concomitant drugs					
	Renin-angiotensin system inhibitor	rasb_prior	rasb_prior	Ras	CM.CMOCCUR.RAS
	Date of first immunosuppressants	—	—	Day	CM.CMSTDTC.PSL
	Prednisolone (yes or no)	IS_bx	fuSteroids_bx	ral steroid p or a	CM.CMOCCUR.PSL
	Immunosuppressants without prednisolone (yes or no)	Non_steroid_IS	—	immuno therapy	CM.CMOCCUR.PSLOTH
Tonsillectomy					
	Tonsillectomy (yes or no)	tonsillectomy	fu_tonsillectomy	Tonsil	SUPPMH.QNAM.OPE
	Date of tonsillectomy	tonsillectomy_dt	—	tonsil date	SUPPMH.QNAM.OPEDATE
Laboratory examinations					
	Date of kidney biopsy	date_bx	date_bx	kidney_biopsy_date	LB.LBDTC.BIOPSY
	Serum creatinine	Creatinine	—	Cr	LB.LBORRES.CREAT
	eGFR[c]	eGFR	gfr_bx_provided	Egfr	LB.LBORRES.EGFR
	Urinary protein (spot)	uprot_bx	uprot	urinprotein1	LB.LBORRES.PROT1
	Urinary protein (24 h)	uprot_24h_bx_provided	uprot_24h	Urinprotein	LB.LBORRES.PROT24
Pathological findings					
	Oxford classification: mesangial hypercellularity (M)	m	m	Oxford1	SUPPMH.QNAM.M
	Oxford classification: endocapillary hypercellularity (E)	e	e	Oxford2	SUPPMH.QNAM.E
	Oxford classification: segmental glomerulosclerosis (S)	s	s	Oxford3	SUPPMH.QNAM.S
	Oxford classification: tubular atrophy/interstitial fibrosis (T)	t	t	Oxford4	SUPPMH.QNAM.T

[a]SDTM: Study Data Tabulation Model.

[b]Not available.

[c]eGFR: estimated glomerular filtration rate.

Discussion

Strength of This Study

Integrating multiple preexisting databases through collaboration between the disease specialist and clinical data manager enabled the use of legacy data. Our project suggested that properly defining CDISC SDTM metadata allowed for the integration of multiple preexisting databases. In this paper, we focused on the technical aspect. Although the utility of this concept has been verified with hypothetical data, there are few reports that generate SDTM data from actual clinical databases focused on technical aspects in detail.

The CDISC SDTM

The definition of metadata using the CDISC SDTM is important. The CDISC is a nonprofit, global organization that consists of pharmaceutical companies, contract research organizations, academic research organizations, and IT vendors. Pharmaceutical companies and contract research organizations account for 34% of the entities within the CDISC, whereas academic research organizations account for only 7% [21]. This imbalance may have arisen because those submitting a regulatory application to the FDA or the Pharmaceuticals and Medical Device Agency are required to comply with CDISC standards [22]. Therefore, there is a strong awareness of the CDISC as a tool for regulatory submissions, but few researchers are aware that the CDISC SDTM concept can be used to standardize data.

The mission of the CDISC is to develop and support global, platform-independent data standards that enable information system interoperability to improve medical research and related areas of health care. Following this statement, we have succeeded in integrating 3 databases by incorporating the CDISC SDTM concept into the standardization of multiple databases. Since this database complies with the standardization of the CDISC SDTM, this integrated database can be compared to other clinical trials or it can be used as a historical control. Our study shows that the CDISC SDTM is not only a necessary tool for applying for the approval of regulatory submissions but also for data standardization and integration. In recent years, the CDISC has partnered with REDCap to make Clinical Data Acquisition Standards Harmonization eCRF metadata available in the REDCap Shared Instrument Library [22,23]. It is expected that the researchers will be able to import CDISC SDTM metadata directly into their REDCap projects for immediate use in clinical trial data collection. In the future, CDISC SDTM data will be generated more easily.

We were able to develop the methodology for integrating multiple preexisting databases in just 6 months. This timely integration was due to the collaboration of specialists in the disease area and a data manager familiar with the CDISC SDTM, allowing each phase to proceed simultaneously and resulting in a fast integration time. Inconsistencies in the coding method of the nominal scale hindered the integration of multiple databases. However, in this study, codes defined individually for each database were automatically recoded to substantially reduce the required work hours, which also contributed to the fast integration time. When coding terms not defined by CT, such as terms that are specific to the disease area, a code list should be created following thorough discussions with specialists, referring to therapeutic area standards [24]. It is important to improve work efficiency by making the best use of existing materials. Although the preexisting databases were integrated in this study, even in cases where the data were updated longitudinally, it is possible to integrate data with the SDTM, provided that the meta-information for the evaluated item is defined.

Currently, there are several medical standards. Observational Medical Outcomes Partnership (OMOP), which is managed by Observational Health Data Science and Informatics [25], aimed to standardize interoperability observational databases such as electric medical records and claim data. HL7 Fast Healthcare Interoperability Resources (FHIR) [26] is the standard for medical information exchange. In this study, we used the CDISC SDTM because, at the time, it was the most widely used standard with many accumulated findings. We plan to expand this project to support the OMOP Common Data Model and FHIR in the future.

Issues for Integration

We observed the following points when integrating the preexisting clinical databases: (1) the variability of the collected items and (2) the complexity of the test code. The items in the preexisting cohort studies used in this project were not standardized and were not defined in detail; therefore, we clarified the meaning of the variables based on expert opinions. Clarifying data definitions is difficult for data managers who lack the requisite background knowledge.

In addition, we were faced with large differences in the number of items collected from each preexisting cohort study. As previously mentioned, 57 and 65 items were collected in cohorts A and B, respectively, far fewer than the 582 items collected in cohort C, which included data related to concomitant medications. However, because information on concomitant medications is often missing, it is considered a difficult item to use for analysis. Generally, information on concomitant medications is not used for analysis and is not collected in precise clinical trials. To avoid complications in the integration process, information collected on concomitant medications should focus on those related to the disease area or should be divided into categories prior to collection. These findings were obtained by scrutinizing the differences in the items collected in each database prior to generating the metadata.

The complexity of the test code was clarified during the generation of the metadata. As described above, the amount of the urinary protein was defined as both “PROT” and “PROT24.” Because the details of proteinuria are not defined in CT, there is a risk for inappropriate metadata. These findings suggest that the generation of metadata requires a deep understanding of the disease in addition to the concepts of the CDISC SDTM. In this study, the clinical data manager who had knowledge of the CDISC SDTM was responsible for generating the metadata in collaboration with a specialist in the disease area. Currently, clinical data managers primarily play an active role in prospective clinical trials. Thus, the main responsibilities of the clinical data manager are planning the clinical trial, assisting

with the creation of the protocol and CRF, cleaning the data, confirming data consistency, and managing data quality in clinical trials. We believe that the clinical data manager will play an important role for data integration projects in the near future. Collaborations between the clinical data manager and the disease specialist will likely become even more important.

Limitations

This project had several limitations. First, the data of the cohort studies did not cover all domains of this disease. In the future, we would like to increase the number of integration examples and generalize the program to cover all domains. Second, a great deal of time was spent manually setting the metadata. In the future, it may be beneficial to automatically refer to the shared metadata from the CDISC Library or to develop a tool that allows artificial intelligence to suggest the metadata using therapeutic area standards. Third, REDCap2SDTM version 2 required input for the ODM format. Several programs that generate ODM or Define-XML data from a spreadsheet are available from the CDISC Open Source Alliance [27]. We will consider embedding these programs into REDCap2SDTM version 2 in the future.

Conclusion

Our results suggest that the CDISC SDTM is useful for integrating multiple preexisting databases with variable names and codes. We hope that this research will contribute to the use of legacy data sets.

Acknowledgments

This work was supported by Grant-in-Aid for Early-Career Scientists (Japan Society for the Promotion of Science [JSPS]) 18K17380; Grant-in-Aid for Scientific Research (C) (JSPS) 19K12867 and 21K10445; Japan Agency for Medical Research and Development (AMED) under grant JP20lk0201061; and Grant-in-Aid for Intractable Renal Diseases Research, Research on Rare and Intractable Diseases, Health and Labour Sciences Research Grants from the Ministry of Health, Labour and Welfare of Japan (grant 20FC1045).

Conflicts of Interest

None declared.

Multimedia Appendix 1
The package of REDCap2SDTM version 2.
[ZIP File, 42 KB - medinform_v11i1e46725_app1.zip]

References

1. Taichman DB, Sahni P, Pinborg A, et al. Data sharing statements for clinical trials: a requirement of the International Committee of Medical Journal Editors. PLoS Med 2017 Jun 5;14(6):e1002315. [doi: 10.1371/journal.pmed.1002315] [Medline: 28582414]
2. Blettner M, Sauerbrei W, Schlehofer B, Scheuchenpflug T, Friedenreich C. Traditional reviews, meta-analyses and pooled analyses in epidemiology. Int J Epidemiol 1999 Feb;28(1):1-9. [doi: 10.1093/ije/28.1.1] [Medline: 10195657]
3. Nichols HB, Schoemaker MJ, Cai J, et al. Breast cancer risk after recent childbirth: a pooled analysis of 15 prospective studies. Ann Intern Med 2019 Jan 1;170(1):22-30. [doi: 10.7326/M18-1323] [Medline: 30534999]
4. CDISC. Clinical Data Interchange Standards Consortium. URL: www.cdisc.org/ [accessed 2023-08-04]
5. SDTM. Clinical Data Interchange Standards Consortium. URL: www.cdisc.org/standards/foundational/sdtm [accessed 2023-08-04]
6. CDISC Submission Data Standards Team. CDISC Study Data Tabulation Model Implementation Guide: Human Clinical Trials Version 3.3: Clinical Data Interchange Standards Consortium; 2018.
7. Define-XML. Clinical Data Interchange Standards Consortium. URL: www.cdisc.org/standards/data-exchange/define-xml [accessed 2023-08-04]
8. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research Electronic Data Capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009 Apr;42(2):377-381. [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
9. Franklin JD, Guidry A, Brinkley JF. A partnership approach for electronic data capture in small-scale clinical trials. J Biomed Inform 2011 Dec;44 Suppl 1(Suppl 1):S103-S108. [doi: 10.1016/j.jbi.2011.05.008] [Medline: 21651992]
10. REDCap. Research Electronic Data Capture. URL: www.project-redcap.org/ [accessed 2023-08-04]
11. Yamamoto K, Ota K, Akiya I, Shintani A. A pragmatic method for transforming clinical research data from the Research Electronic Data Capture “REDCap” to Clinical Data Interchange Standards Consortium (CDISC) Study Data Tabulation Model (SDTM): development and evaluation of REDCap2SDTM. J Biomed Inform 2017 Jun;70:65-76. [doi: 10.1016/j.jbi.2017.05.003] [Medline: 28487263]

12. Yamamoto K. Introduction to Research Electronic Data Capture (REDCap) and REDCap2SDTM, a conversion tool to facilitate clinical research data sharing. Article in Japanese. Brain Nerve 2017 Jul;69(7):848-855. [doi: 10.11477/mf.1416200830] [Medline: 28740000]
13. Oda T, Chiu SW, Yamaguchi T. Semi-automated conversion of clinical trial legacy data into CDISC SDTM standards format using supervised machine learning. Methods Inf Med 2021 May;60(1-02):49-61. [doi: 10.1055/s-0041-1731388] [Medline: 34237784]
14. Cholesterol Treatment Trialists' Collaboration. Harmonisation of large-scale, heterogeneous individual participant adverse event data from randomised trials of statin therapy. Clin Trials 2022 Dec;19(6):593-604. [doi: 10.1177/17407745221105509] [Medline: 35815805]
15. Takahara S, Saito TI, Imai Y, Kawakami T, Murayama T. A use-case analysis of Clinical Data Interchange Standards Consortium/Study Data Tabulation Model in academia in an investigator-initiated clinical trial. Nagoya J Med Sci 2022 Feb;84(1):120-132. [doi: 10.18999/nagjms.84.1.120] [Medline: 35392016]
16. Koyama A, Igarashi M, Kobayashi M, Research Group on Progressive Renal Diseases. Natural history and risk factors for immunoglobulin A nephropathy in Japan. Am J Kidney Dis 1997 Apr;29(4):526-532. [doi: 10.1016/S0272-6386(97)90333-4] [Medline: 9100040]
17. Controlled terminology. Clinical Data Interchange Standards Consortium. URL: www.cdisc.org/standards/terminology/controlled-terminology [accessed 2023-08-04]
18. ODM. Clinical Data Interchange Standards Consortium. URL: www.cdisc.org/standards/data-exchange/odm [accessed 2023-08-04]
19. Tomino Y, Sakai H, Special Study Group (IgA Nephropathy) on Progressive Glomerular Disease. Clinical guidelines for immunoglobulin A (IgA) nephropathy in Japan, second version. Clin Exp Nephrol 2003 Jun;7(2):93-97. [doi: 10.1007/s10157-003-0232-4] [Medline: 14586726]
20. Working Group of the International IgA Nephropathy Network and the Renal Pathology Society, Roberts ISD, Cook HT, et al. The Oxford classification of IgA nephropathy: pathology definitions, correlations, and reproducibility. Kidney Int 2009 Sep;76(5):546-556. [doi: 10.1038/ki.2009.168] [Medline: 19571790]
21. Membership. Clinical Data Interchange Standards Consortium. URL: www.cdisc.org/membership [accessed 2023-08-04]
22. Providing regulatory submissions in electronic format -- standardized study data. US Food & Drug Administration. 2021 Jun. URL: www.fda.gov/regulatory-information/search-fda-guidance-documents/providing-regulatory-submissions-electronic-format-standardized-study-data [accessed 2023-11-09]
23. eCRF portal. Clinical Data Interchange Standards Consortium. URL: www.cdisc.org/kb/ecrf [accessed 2023-08-04]
24. Therapeutic areas. Clinical Data Interchange Standards Consortium. URL: www.cdisc.org/standards/therapeutic-areas [accessed 2023-08-04]
25. OHDSI. Observational Health Data Sciences and Informatics. URL: www.ohdsi.org/ [accessed 2023-08-04]
26. Enabling health interoperability through FHIR. HL7 FHIR Foundation. URL: https://fhir.org/ [accessed 2023-08-04]
27. COSA repository directory. CDISC Open Source Alliance. URL: https://cosa.cdisc.org/ [accessed 2023-08-04]

Abbreviations

CDISC: Clinical Data Interchange Standards Consortium
CRF: case report form
CT: controlled terminology
FDA: Food and Drug Administration
FHIR: Fast Healthcare Interoperability Resources
ICMJE: International Committee of Medical Journal Editors
IgA: immunoglobulin A
LOINC: Logical Observation Identifiers Names and Codes
ODM: Operational Data Model
OMOP: Observational Medical Outcomes Partnership
REDCap: Research Electronic Data Capture
SDTM: Study Data Tabulation Model
SNOMED: Systematized Nomenclature of Medicine

Edited by J Klann; submitted 11.03.23; peer-reviewed by A Gao, A Loban, S Hume; revised version received 13.09.23; accepted 14.09.23; published 21.12.23.

Please cite as:
Matsuzaki K, Kitayama M, Yamamoto K, Aida R, Imai T, Ishida M, Katafuchi R, Kawamura T, Yokoo T, Narita I, Suzuki Y
A Pragmatic Method to Integrate Data From Preexisting Cohort Studies Using the Clinical Data Interchange Standards Consortium (CDISC) Study Data Tabulation Model: Case Study
JMIR Med Inform 2023;11:e46725
URL: https://medinform.jmir.org/2023/1/e46725
doi:10.2196/46725

Original Paper

Assessment and Improvement of Drug Data Structuredness From Electronic Health Records: Algorithm Development and Validation

Ines Reinecke[1], BA, MA; Joscha Siebel[1]; Saskia Fuhrmann[2,3], Dr Rer Nat; Andreas Fischer[3], MA; Martin Sedlmayr[2], Prof Dr, Dr Rer Nat; Jens Weidner[1], MA; Franziska Bathelt[1], Dr Rer Nat

[1]Institute for Medical Informatics and Biometry, Carl Gustav Carus Faculty of Medicine, Technische Universität Dresden, Dresden, Germany

[2]Center for Evidence-Based Healthcare, Carl Gustav Carus Faculty of Medicine, Technische Universität Dresden, Dresden, Germany

[3]Hospital Pharmacy, University Hospital Carl Gustav Carus, Dresden, Germany

Corresponding Author:
Ines Reinecke, BA, MA
Institute for Medical Informatics and Biometry
Carl Gustav Carus Faculty of Medicine
Technische Universität Dresden
Fetscherstraße 74
Dresden, 01307
Germany
Phone: 49 35145887 ext 12975
Email: ines.reinecke@uniklinikum-dresden.de

Abstract

Background: Digitization offers a multitude of opportunities to gain insights into current diagnostics and therapies from retrospective data. In this context, real-world data and their accessibility are of increasing importance to support unbiased and reliable research on big data. However, routinely collected data are not readily usable for research owing to the unstructured nature of health care systems and a lack of interoperability between these systems. This challenge is evident in drug data.

Objective: This study aimed to present an approach that identifies and increases the structuredness of drug data while ensuring standardization according to Anatomical Therapeutic Chemical (ATC) classification.

Methods: Our approach was based on available drug prescriptions and a drug catalog and consisted of 4 steps. First, we performed an initial analysis of the structuredness of local drug data to define a point of comparison for the effectiveness of the overall approach. Second, we applied 3 algorithms to unstructured data that translated text into ATC codes based on string comparisons in terms of ingredients and product names and performed similarity comparisons based on Levenshtein distance. Third, we validated the results of the 3 algorithms with expert knowledge based on the 1000 most frequently used prescription texts. Fourth, we performed a final validation to determine the increased degree of structuredness.

Results: Initially, 47.73% (n=843,980) of 1,768,153 drug prescriptions were classified as structured. With the application of the 3 algorithms, we were able to increase the degree of structuredness to 85.18% (n=1,506,059) based on the 1000 most frequent medication prescriptions. In this regard, the combination of algorithms 1, 2, and 3 resulted in a correctness level of 100% (with 57,264 ATC codes identified), algorithms 1 and 3 resulted in 99.6% (with 152,404 codes identified), and algorithms 1 and 2 resulted in 95.9% (with 39,472 codes identified).

Conclusions: As shown in the first analysis steps of our approach, the availability of a product catalog to select during the documentation process is not sufficient to generate structured data. Our 4-step approach reduces the problems and reliably increases the structuredness automatically. Similarity matching shows promising results, particularly for entries with no connection to a product catalog. However, further enhancement of the correctness of such a similarity matching algorithm needs to be investigated in future work.

(JMIR Med Inform 2023;11:e40312) doi:10.2196/40312

KEYWORDS

secondary usage; Observational Medical Outcomes Partnership; OMOP; drug data; data quality; Anatomical Therapeutic Chemical; ATC; RxNorm; interoperability

Introduction

Background

Over the last decade, the amount of electronically available data in the health care domain has increased enormously worldwide. Much of the data is generated during the processing of administrative claims, through documentation processes in electronic health records (EHRs) performed during patient treatments, or via data feeds from mobile devices providing patient-reported outcomes. Therefore, it is not surprising that real-world data (RWD) are becoming more important for health care research. RWD studies can be considered complementary to randomized controlled trials (RCTs), as they allow the results of RCTs to be confirmed in much larger cohorts and over a longer period. Compared with RCTs, RWD studies allow for better external validity and better generalizability, and they not only offer opportunities for long-term surveillance of drug products but also are cost-effective and time-saving [1].

Drug surveillance systems, such as the US Food and Drug Administration's Sentinel initiative, are critical for promoting postmarket drug safety [2-8]. The European Medicines Agency has also started to establish research infrastructure based on RWD to support pharmacovigilance [9]. In addition, the European Health Data and Evidence Network [10] emerged to establish transnational research networks based on a common data model that enables standardized RWD and methods for observational studies to generate real-world evidence. Recently, the European Health Data and Evidence Network has begun to collaborate with the European Medicines Agency to address COVID-19 [11].

However, the original purpose of RWD generation during patient treatment is not primarily aimed at its use in research. Therefore, notable problems have been identified regarding the replication and validity of observational research results based on RWD. To ensure the reliability and robustness of the results from RWD research, these issues have to be addressed, as they become even more important when observational studies are conducted across countries at large scale.

Data harmonization, the use of international standards and terminologies, a common data model, methods, and tools for data analyses that increase the reproducibility of results are needed [12]. These gaps are being addressed by the International Observational Health Data Sciences and Informatics community, which provides the common data model called the Observational Medical Outcomes Partnership (OMOP) and standardized analysis tools based on OMOP. It also includes standardized vocabularies that contain translations between national terminologies and internationally acknowledged terminologies, for example, Systematized Nomenclature of Medicine-Clinical Terms, Logical Observation Identifiers Names and Codes, Anatomical Therapeutic Chemical (ATC) classification, and RxNorm [13]. OMOP allows RWD to be stored in the same way, regardless of data origin, thus ensuring the use of RWD in international, large-scale observational studies. Compared with similar projects such as Informatics for Integration Biology and the Bedside or the National Patient-Centered Clinical Research Network, Observational Health Data Sciences and Informatics-OMOP meets the needs of observational RWD studies well [14]. Many RWD studies on OMOP have shown that drug data at the ingredient level are sufficient to answer their research questions [15]. Although drug data with details on dosage and units for drug exposure can be important for observational research on drug effectiveness and drug safety with the same drug at different doses, the availability of the drug ingredient is the least common denominator and the basic requirement for drug-related RWD studies on OMOP. Therefore, drug prescription data must be available in a structured format that does not necessarily include the name of the drug product but at least the ingredient information. For drug utilization research, the World Health Organization recommends the use of ATC classification, which divides drugs into different groups based on the organ or system on which they act [16]. ATC classification includes a hierarchy based on 5 different levels, with ATC level 5 being the chemical substance that represents the active ingredient of a drug product [17]. Each approved drug product on the market is assigned a specific ATC level 5 code. The National Institutes of Health Collaboratory recommends assessing and reporting the quality of EHR data for clinical reuse in terms of data completeness, accuracy, and consistency [18]. Weiskopf et al [19] also determined that the completeness and correctness of data are of special importance for data quality improvement.

Objective

To the best of our knowledge, there is no existing approach to systematically analyze and improve the structuredness of drug prescription data for observational research. Thus, in this study, we systematically analyzed the structuredness of EHR drug prescriptions to determine the ratio between structured drug prescription data containing ATC code level 5 and free-text drug prescriptions without an available standard concept based on the 14 ATC groups of level 1. In addition, we presented an approach to improve the structuredness of drug prescription data by introducing an automatic detection method for ATC code determination. To ensure the robustness and accuracy of the results of automatic detection, we introduced a validation step based on existing text-mining algorithms.

Methods

Study Details

This retrospective, noninterventional study systematically reviewed drug prescriptions based on real-world observational data at the University Hospital Carl Gustav Carus Dresden (UKD), Germany. This study was based on fully anonymized data and did not include any correlations with individual patients. All inpatient drug prescriptions, including acute medications, from 2016 to 2020 were included in the study, without restriction to specific conditions or treatments. The original data were recorded in the ORBIS hospital information system from Dedalus, using the ORBIS module, "KURV," that represents the patient curve including medication data. A total of 1,768,153 drug prescriptions were reviewed from the hospital information system records. Drug prescription data from other systems (eg, intensive care units and chemotherapy) were excluded because data in those systems were completely

structured and stored in separate backend systems. The data used for this study were provided by the Data Integration Centre at the UKD, which was established with funding from the German Federal Ministry of Education and Research as part of the Medical Informatics Initiative in Germany.

Ethics Approval

The study was approved by the Ethics Committee of the Technical University of Dresden as a retrospective, observational, noninterventional, nonhuman subject study (SR-EK-521112021).

Data Set Details

The following 2 data sets were used: drug prescription data (data set 1) and drug product catalog data (data set 2). The drug product catalog was exported from the UKD Enterprise-Resource-Planning system on November 16, 2021, and contained the drug product name, drug ingredient name, ATC level 5 code, drug dose and unit information, and legacy products. In addition, 2 other data sets were derived from data set 1. First, an aggregated data set (data set 3) was generated based on the grouped data set 1 for all the unstructured drug prescription entries. The grouping activity to create data set 3 was performed on the MEDICATION column of data set 1 by grouping all entries in the data element MEDICATION using the Python library Pandas and its *groupby* function. Frequency information for each unique MEDICATION record was added to data set 3. The data set (data set 4) contained a subset of the first 1000 most frequent entries from data set 3 and additional results from the manual evaluation step.

All the metadata elements of the data sets presented above that are relevant to this study are illustrated and described in detail in Table 1. Drug prescriptions selected from the drug catalog data are labeled as *structured data* (eg, “IBUPROFEN STADA 600 mg Zäpfchen | [Ibuprofen natrium, Ibuprofen]”) based on the contents of the STRUCTURE column of data set 1. Drug prescriptions that were not selected from the drug product catalog are designated as *unstructured data* (eg, “Ibuprofen 600” and “Ibuprofen”).

Table 1. Description of relevant data set with its metadata elements.

Data set and data element	Data type	Description
DS[a]1 initial data set with all drug prescriptions		
MEDICATION	String	Free text or predefined value, chosen from an available fixed drop-down menu that contains product names, derived from the drug product catalog, when creating a new drug prescription
YEAR	Number	Extracted from the prescription start date information for further statistical analyses
STRUCTURE	Boolean	TRUE if MEDICATION was chosen from the drug catalog or FALSE if the free text was entered
ATC[b]_L5	String	ATC code level 5 available in case STRUCTURE is true otherwise empty
DS2 drug catalog data		
Product_name	String	Product name as listed in the ERP[c] system
Ingredient_name	String	Ingredient name as listed for the product
Atc_code	String	ATC code level 5
DS3 grouped DS1 by MEDICATION data element		
MEDICATION	String	Grouped unstructured free-text entries
FREQUENCY	Number	Summed up the occurrence of the MEDICATION text field to determine the most relevant free-text drug prescriptions
Step1	String	Algorithm 1 result as an ATC code or empty if no match
Step2	String	Algorithm 2 result as an ATC code or empty if no match
Step3	String	Algorithm 3 result as an ATC code or empty if no match
DS4 most frequent 1000 entries of DS3 (sorted by frequency)		
MEDICATION	String	Grouped unstructured free-text entries
FREQUENCY	Number	Summed up the occurrence of the medication text field
Step1	String	Algorithm 1 result as an ATC code or empty if no match
Step2	String	Algorithm 2 result as an ATC code or empty if no match
Step3	String	Algorithm 3 result as an ATC code
Eval1	Boolean	Algorithm 1 evaluation result
Eval2	Boolean	Algorithm 2 evaluation result
Eval3	Boolean	Algorithm 3 evaluation result
True12	Boolean	TRUE if the same result for algorithm 1+2
True13	Boolean	TRUE if the same result for algorithm 1+3
True23	Boolean	TRUE if the same result for algorithm 2+3
True123	Boolean	TRUE if the same result for algorithm 1+2+3
CORRECT	String	Corrected ATC code. in case no algorithm determined the correct result, entered manually in the evaluation step
COMMENTS	String	Any comments or additional information if needed
FINAL	String	Finally determined ATC code for all entries or labels in case no ATC code could be determined (labels are introduced in the methods validation section in detail)

[a]DS: data set.

[b]ATC: Anatomical Therapeutic Chemical.

[c]ERP: Enterprise-Resource-Planning.

Data Analysis

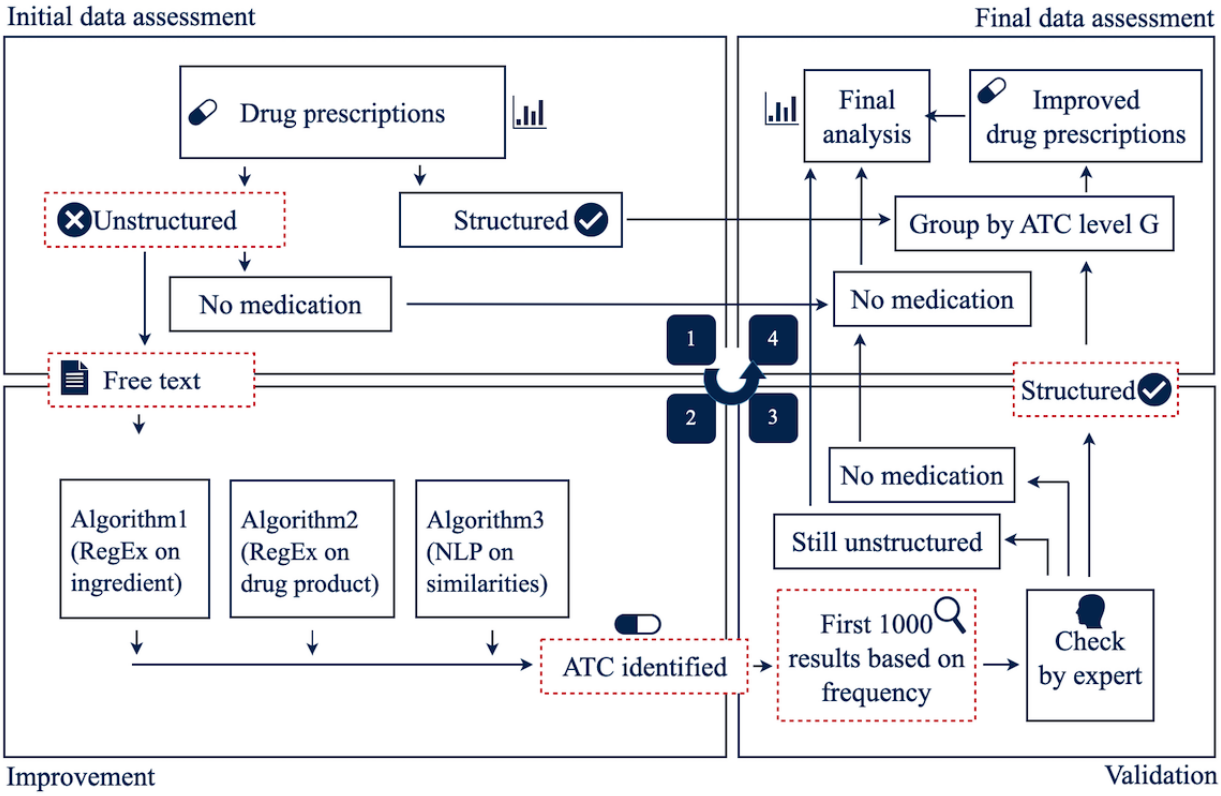
Overview

Data analysis consisted of a 4-step process, as shown in Figure 1. The first step of the process was an initial data quality analysis to determine the overall ratio of structured to unstructured drug prescriptions.

To improve the structure of the drug prescriptions, 3 existing algorithms were applied to automatically identify correct ATC codes for the unstructured drug prescriptions. The identified ATC codes were then manually reviewed by experts

(pharmacists and medical information scientists) and checked for correctness. This step also included the identification of existing patterns that can help conclude the reliability of the automatically identified ATC codes for unstructured data. Finally, the results of the previous 3 steps were consolidated to assess the degree of improvement achieved in unstructured drug prescriptions. To ensure expert coverage of the entire process, an interdisciplinary team of pharmacists, computer scientists, and medical informatics researchers was formed.

Figure 1. Data analysis 4-step approach. ATC: Anatomical Therapeutic Chemical; NLP: natural language processing.



Initial Data Assessment

Initially, the ratio of structured to unstructured drug prescriptions was determined for data set 1. For this purpose, the STRUCTURE data element of data set 1 was used to subdivide the data into 2 groups. If the value of STRUCTURE was TRUE, the record was considered to be structured; otherwise, it was unstructured. Subsequently, the unstructured subset of drug prescriptions was grouped by the data element MEDICATION as data set 3, and the frequency was calculated and added as the data element FREQUENCY.

The first manual review of the grouped drug prescriptions (data set 3) was done by the interdisciplinary team of experts to identify records that are not drug prescriptions but other instructions, such as orders for blood counts or other laboratory and measurement orders (eg, “BGA”—laboratory request for blood gas analysis, “BE”—request to nurses for taking a blood sample, and “BB”—laboratory request for blood count). This task resulted in a set of rules (Multimedia Appendix 1) to allow the automated search and identification of medication entries that needed to be excluded for further steps.

Improvement

Unstructured drug prescriptions (usually provided as free text) were used as inputs for the improvement step. Preprocessing of the drug prescription was not performed previously. In this step, 3 different algorithms were implemented to automatically identify ATC level 5 codes based on the MEDICATION text.

The algorithms were based on a different mechanism for matching the MEDICATION text of data set 1 with the product catalog data elements INGREDIENT_NAME and PRODUCT_NAME of data set 2, as described in detail in Table 2.

Table 2. An overview of algorithms for Anatomical Therapeutic Chemical (ATC) code identification for unstructured drug prescriptions.

Algorithm	Mechanism	Data input for comparison		Result data
		Data set 1	Data set 2	
1	String comparison	MEDICATION	Ingredient_name	ATC code
2	String comparison	MEDICATION	Product_name	ATC code
3	Similarity matching	MEDICATION	INGREDIENT_NAME and PRODUCT_NAME	ATC code+similarity score

Algorithms 1 and 2 rely on simple string comparisons to recognize either the ingredient name or product name within the drug prescription. Algorithm 3 performs natural language processing (NLP) based on similarity matching between the data element MEDICATION in data set 1 and the 2 data elements PRODUCT_NAME and INGREDIENT in data set 2 with the Python library *FuzzyWuzzy* [20] using Levenshtein distance because it has shown promising results in other health care research areas [21,22]. The best similarity score result was 100, which meant that the components of the string MEDICATION were entirely contained in INGREDIENT_NAME or PRODUCT_NAME. The lower the similarity score, the less similar the MEDICATION string is compared with the drug catalog entries. This algorithm provided up to 3 possible ATC codes, sorted in descending order based on their similarity scores. To determine the most promising method of the *FuzzyWuzzy* library for our implementation, we defined that the word order in the data element MEDICATION is irrelevant and can be different from the compared strings in INGREDIENT_NAME and PRODUCT_NAME. All words from the entry of the data element MEDICATION must be included in the entry of INGREDIENT_NAM*E* or PRODUCT_NAME, but not vice versa. This led to the implementation of the method *token_set_ratio*. This method tokenizes both strings to be compared, changes the upper case to the lower case, and removes punctuation. It then sorts the tokens alphabetically and split them into 2 groups: the intersection group (tokens that are the same in both strings) and the remainder group (tokens that differ in compared strings). The *token_set_ratio* method compares the intersection group with the intersection and remainder of the first string and then the same with the remainder of the other string and finally takes the highest results of this comparison as the final result. As shown in the following example (Textbox 1), the *token_set_ratio* method provides the best results concerning the given requirements.

Textbox 1. An example of the token_set_ratio method.

```
d1 = “Stada paracetamol”
d2 = “paracetamol Stada 400 mg”
Print(“Ratio: ”, fuzz.ratio(d1.lower(),d2.lower()))
Print(“Partial Ratio: ”, fuzz.partial_ratio(d1.lower(),d2.lower()))
Print(“Token Sort Ratio: ”, fuzz.token_sort_ratio(d1.lower(),d2.lower()))
Print(“Token Set Ratio: ”, fuzz.token_set_ratio(d1.lower(),d2.lower()))
Ratio: 54
Partial Ratio: 65
Token Sort Ratio: 83
Token Set Ratio: 100
```

The algorithms were applied to data sets 1 and 3. The results of the algorithms in data set 3 were also used in data set 4. The concordance between the results for each permutation (algorithms 1+2, 1+3, 2+3, and 1+2+3) was also calculated. The complete source can be accessed on Zenodo [23].

Validation

The validation step consisted of manual checks of the automatically generated ATC codes by the same interdisciplinary team as in the previous steps. It was performed on a subset of the most common free-text prescriptions. To maintain the validation effort proportionate to the benefit, a minimum target was defined for the manual validation process of unstructured drug prescriptions to cover at least 80% of structured and manually validated unstructured entries combined. During the validation step, information was added to each algorithm to determine whether the correct ATC code, wrong ATC code, or no ATC code was identified. If no algorithm identified the correct ATC code, it was determined by manual validation when possible. If an entry was found to generally have no drug prescription, it was marked as an additional entry without drug prescription with the keyword “nomed.” For drug prescriptions that require further specification to determine the exact ATC level 5 code, the manual review checks whether the ATC level 4 or 3 code can be determined based on the free text of the drug prescription, otherwise the entry was flagged as unspecific with the keyword “unspec.” All unstructured drug prescription entries for which no validation of the automatically generated ATC codes was performed were marked with the keyword “no_eval.”

The results of the manual validation were summarized to identify any patterns that can help improve the robustness of the results of automatically detected ATC codes based on the total findings and correctness of each algorithm, the concordance level between the results of algorithms 1, 2, and 3, and the Levenshtein similarity score for algorithm 3. For algorithm 3, we used a 2-tailed *t* test implemented in Python to determine whether there was a significant difference between the means of the Levenshtein similarity score for the correct and incorrect results.

In addition, the incorrect results were examined in more detail by the interdisciplinary team to identify patterns that would reveal important reasons and similarities related to the ingredients of concern (ATC) to the greatest extent possible.

Final Data Assessment

For the final data assessment, the results of the step improvement and validation documented in data set 4, including correctly identified ATC level 5 codes or “nomed,” “unspec,” or

"no_eval" labels, were merged with the original drug prescription data from data set 1. Thus, the final data assessment was executed based on the algorithm results and manual validation. The total number of drug prescription records was determined for each of the 14 ATC groups, including the proportion of structured versus unstructured data per ATC group. In addition, the total number of unique ATC level 5 codes, including their structuredness, as well as the most frequent ATC level 5 codes used in drug prescription of the data set 1 are presented. This allows a ranking of the structuredness based on the ATC groups and ATC codes.

Results

Initial Data Assessment

The initial assessment revealed 843,980 (n=1,768,153, 47.73% drug prescriptions in the data set 1) structured drug prescriptions. The proportion of unstructured drug prescriptions that required further investigation was 52.27% (924,173 drug prescription entries). A small set of rules, for example, all drug prescription entries starting with laboratory or measurement orders (Multimedia Appendix 1), identified a total of 160,896 (9.1% of all drug prescriptions) entries as no drug prescription data and reduced the unstructured drug prescriptions requiring review for the next steps to 763,277 (43.17% of all drug prescriptions).

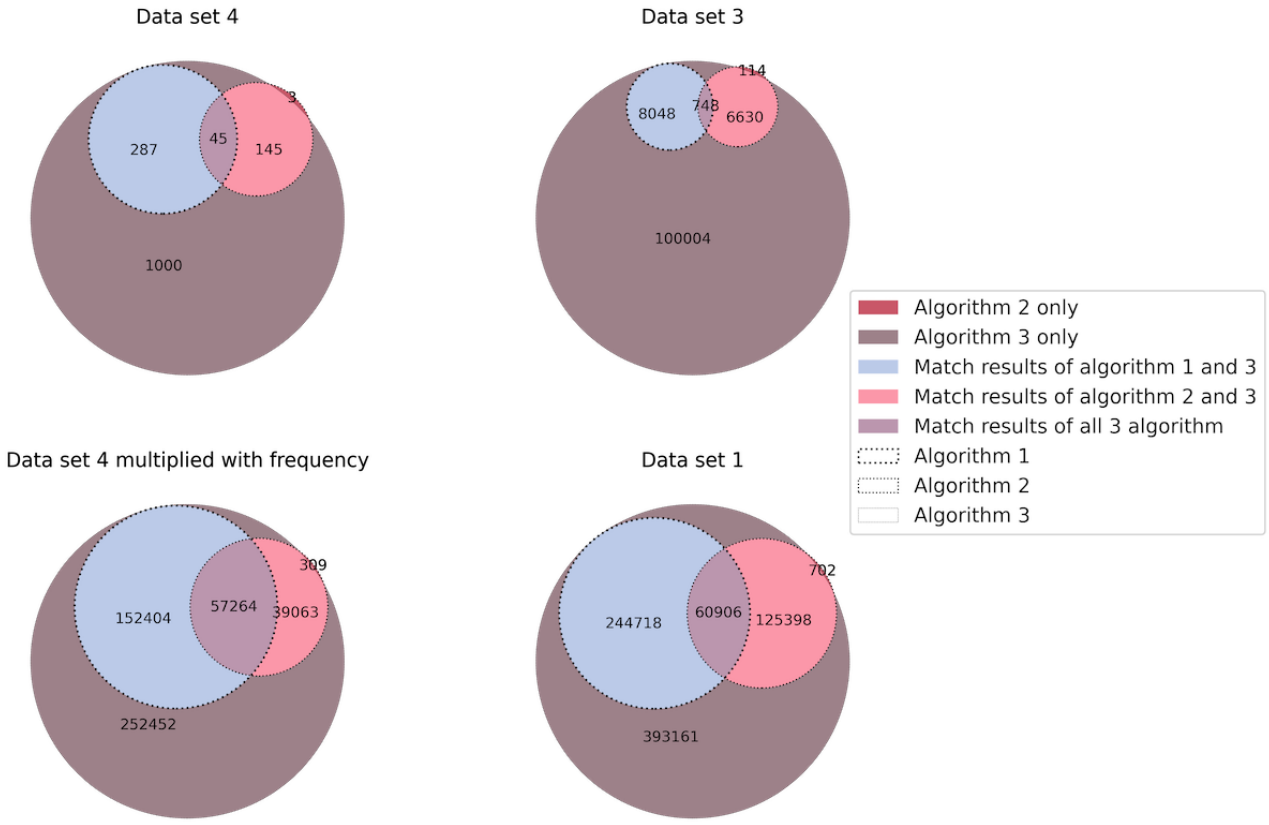
Grouping the unstructured drug prescriptions based on the MEDICATION data element of data set 1 resulted in a total of 100,004 (n=924,173, 10.82%) unique free-text entries that were entered as medication prescription information and stored as data set 3 after adding the frequency for each free text.

Improvement

The quantitative performance of the algorithms was very different, as each algorithm returned a different number of results. Algorithm 3 provides an ATC code for all unstructured drug prescriptions, owing to its implementation and nature.

Algorithm 1 (based on ingredient matching) identified ATC codes for 8048 unique free texts. Multiplied by the frequency of each text entry, this yielded a total of 244,718 (32.06%) drug prescriptions of the total 763,277 unstructured drug prescriptions. The quantitative outcome performance of algorithm 2 (based on the drug product) is lower than that of algorithm 1, as it identified ATC codes for 6744 unique free texts. This represents a total of 126,100 (16.52%) drug prescriptions of the total 763,277 unstructured drug prescriptions. At this point, no statement can be made about the correctness of the algorithm results, but the analysis of the match rate between all algorithms shows matching rates for the total number of unstructured drug prescriptions and the most frequent 1000 free-text entries as illustrated in Figure 2.

Figure 2. Match rates of algorithm results calculated for all data sets under inspection.



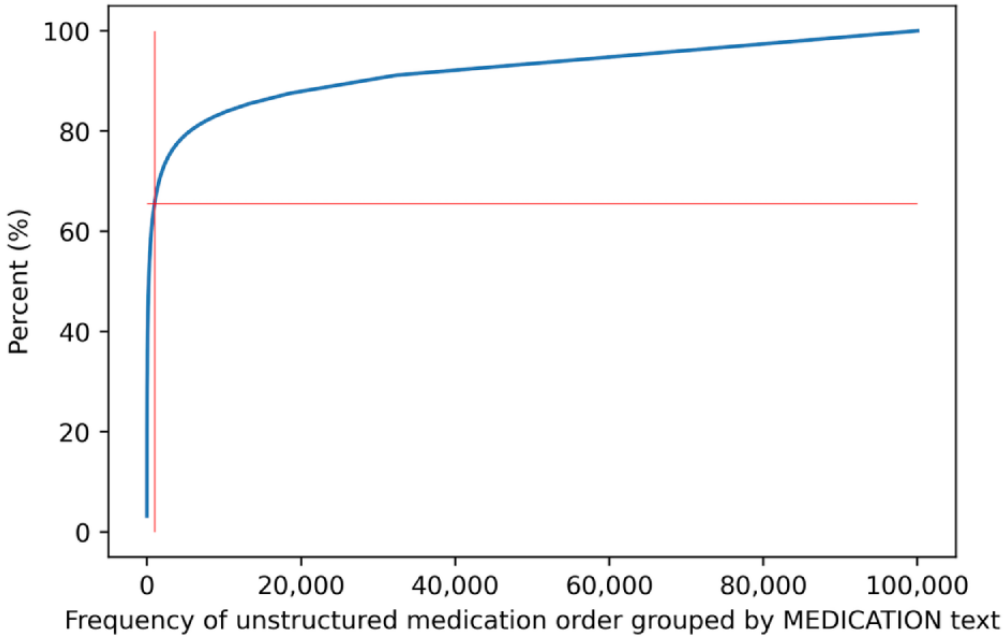
Validation

The manual validation step was performed on the most frequent 1000 free-text entries, which already covered 66.56% (615,129/924,173) of all unstructured drug prescriptions, as shown in Figure 3. Together with the proportion of structured drug prescriptions (843,980/1,768,153, 47.73%) and entries without medication (166,307/1,768,153, 9.4%) that were

identified during the initial data analysis, the structuredness could potentially be increased to 85.18% (1,506,059/1,768,153) of all medication prescriptions.

For the most frequent 1000 free-text entries (data set 4), algorithm 1 returned 286 (28.6%) correct results, 1 (0.1%) incorrect result, and no results for 713 (71.3%) entries. Algorithm 2 returned 142 (14.2%) correct results, 6 (0.6%) incorrect results, and no results for 852 (85.2%) unique entries. Algorithm 3 returned 765 (76.5%) correct results and 235 (23.5%) incorrect results. We also determined the correctness in terms of the result match rates between the algorithms, as shown in Figure 2, for data set 4. After the manual validation of data set 4, the returned ATC codes were always correct if all algorithms or algorithms 1 and 2 returned the same results.

Figure 3. Percentage of the most frequent 1000 of all unstructured drug prescriptions.



For the matching results of algorithms 1 and 3, we noted a minor discrepancy and identified 5 incorrect results out of 286, which were related to sodium chloride drug prescriptions in 4 out of 5 cases, and another incorrect result was returned for the ingredient aciclovir. The manual review revealed that an ATC code could not be provided because of missing details that was due to the ATC code at ATC level 1 varying by route of administration (eg, oral, parenteral, and conjunctival). For the matching results of algorithms 2 and 3, we identified only 1 incorrect result related to the ingredient telmisartan because the drug prescription was for a combination drug (telmisartan and diuretics), whereas the algorithms identified only the ATC code of the single ingredient telmisartan.

For the data set 4, a significant difference in the means of the Levenshtein similarity score was found between the correct and incorrect results (see Table 3 for descriptive statistics) with a P value of 2.4×10^{-47}, which is well below the significance level α (.05). This means that the higher the Levenshtein similarity score, the higher the probability of result correctness. Speaking in terms of absolute numbers, for entries with a Levenshtein similarity score >84.28, the results can be assumed to be correct with a low error rate.

Table 3. Descriptive statistics for Levenshtein similarity score for correct and wrong results.

Descriptive statistics	Algorithm 3	
	Correct	Wrong
Count	766	234
Frequency, n (%)	416,585	84,598
Values, mean (SD)	84.28 (14.86)	67.18 (15.52)
Range (%)	21-100	29-100
Percentile		
25th percentile	76	55
50th percentile	87	63
75th percentile	96	75

The 234 incorrect results returned only by algorithm 3, with 4.78% (84,598/1,768,153) of drug prescriptions, can be categorized into four groups: (1) manually identified additional entries without drug prescriptions for which the applied rules did not work; (2) specification generally not possible owing to missing information; (3) restriction to ATC level 3 or 4 because of nonspecific drug prescription information; and (4) other reasons. We found 16 entries (multiplied by frequency=5411) with no additional drug prescription entries. For an additional 11 (multiplied by frequency=2187) entries, no ATC code could be provided because the dosage form or dose was missing. For 2 drug prescriptions (multiplied by frequency=2610) of insulin therapies, there was a restriction to ATC level 3. Another 2 (multiplied by frequency=887) entries were restricted to ATC level 4 sodium chloride prescriptions.

For the other 203 misidentified entries, we examined the subset of 26 entries where algorithm 3 returned results with a Levenshtein similarity value of ≥80 because it is an indication of correctness but unfortunately did not apply to all results. A small group of 26 results with a Levenshtein similarity value of ≥80 was still incorrect. The main reason for the errors in these results was that the ATC codes for the ingredients differed by dosage form and when combined in a drug product, as shown in Table 4. Most incorrect results (15 out of 26) were caused by the absence of the ingredient dosage form in the free text, especially for sodium chloride, prednisolone, dimetindene, aciclovir, and hydrocortisone. The full data set 4 with all the algorithm outcome quality data elements listed in Table 1 is available in Multimedia Appendix 2.

Table 4. Wrong results of algorithm 3 with Levenshtein similarity score ≥80.

Medication free text	Wrong result	Levenshtein similarity score	Correct result	Reason
ASS RATIOPHARM 100 mg TAH Tabletten|(Acetylsalicylsäure)	N02BA01	89	B01AC06	Similarity of words
Prednisolon	S01CA53	100	H02AB06	Dosage form
MAGNESIUM VERLA 300 Orange Granulat|(Magnesium-Ion)	A12CC05	100	V06XX02	Similarity of words
ARILIN 500 Filmtabletten | (Metronidazol)	G01AF01	100	P01AB01	Similarity of words
CANDESARTAN HEXAL comp 16 mg/12.5 mg Tabletten | (Candesartan)	C09CA06	89	C09DA26	Combination product
Heparin	C05BA03	100	B01AB01	Dosage form
PREDNISOLON	S01CA53	100	H02AB06	Dosage form
FENISTIL Injektionslösung | (Dimetinden)	D04AA13	100	R06AB03	Dosage form
ACIC 250 PI Via Pulver z.Herst.e.Infusionslösg. | (Aciclovir)	D06BB03	100	J05AB01	Dosage form
NaCl 0.9%	B05CB01	100	B05BB11	Dosage form
VALSARTAN HEXAL comp.160mg/12,5mg Filmtabletten | (valsartan)	C09CA03		C09DA23	Combination product
Prednisolon mg	S01CA53	88	H02AB06	Dosage form
NaCL 0.9%	B05CB01	100	B05BB11	Dosage form
ACIC 200 Tabletten | (Aciclovir)	D06BB03	100	J05AB01	Dosage form
ACIC 500 PI Via Pulver z.Herst.e.Infusionslösg. | (Aciclovir)	D06BB03	100	J05AB01	Dosage form
Simvastatin	C10BA02	100	C10AA01	No combination product
CANDESARTAN HEXAL comp 8 mg/12.5 mg Tabletten | (Candesartan)	C09CA06	89	C09DA26	Combination product
NaCL 0.9% | (Natrium-Ion, Chlorid)	B05CB01	100	B05BB11	Dosage form
C) FENISTIL 1 Ampulle als Bolus | (Dimetinden)	D04AA13	100	R06AB03	Dosage form
HCT	C09DX01	100	C03AA03	Shortness of text
Allopurinol	M04AA51	100	M04AA01	Combination product
Prednisolon 5 mg	S01CA53	81	H02AB06	Dosage form
HYDROCORTISON 10 mg Jenapharm Tabletten | (Hydrocortison)	S01BA02	81	H02AB09	Dosage form
Simvastatin 20 mg	C10BA02	100	C10AA01	No combination
NaCl 0.9%	B05CB01	100	B05BB11	Dosage form

Final Data Assessment

Compared with the initial data assessment in which we could only distinguish between structured and unstructured drug prescriptions, we were able to perform a percentage distribution between structured and unstructured drug prescriptions for each of the 14 ATC level 1 groups after applying the algorithm. The final results are presented in Table 5, which shows the number of structured drug prescriptions versus the number of unstructured drug prescriptions per ATC level 1 group. The total number of drug prescriptions per ATC level 1 group, including percentages, provides an overview of the most and least frequently prescribed drugs sorted by the 14 ATC level 1 groups. For completeness, we added 3 additional rows to Table 5 containing the number of unstructured entries identified as other orders (no_med), the number of unstructured entries identified as unspecified entries (unspec), and the remaining unstructured data for which no validation was performed; thus, no statement on the correct ATC code was possible. ATC level 1 group "N – Nervous system" was the most common group with 24.1% (322,286/1,337,565) of the initial data set 1, followed by "B – Blood and blood forming organs," "A – Alimentary tract and metabolism," and "C – Cardiovascular system" with approximately 19% each.

Figure 4 illustrates the structuredness of the data for each of the 14 ATC level 1 groups. The ATC level 1 group with the most structured data was "S – Sensory organs" with 98.03% (5077/5179) structured data, followed by the group "H – Systemic hormonal drugs, excluding sex hormones and insulins" with 79.9% (51,296/64,199) structured data. ATC level 1 groups "R – Respiratory system," "C – Cardiovascular system," "J – Anti-infective for systemic use," "V – Various," "B – Blood and blood forming organs," and "N – Nervous system," ranged from 61% to 70% structured data. ATC group "P – Antiparasitic products, insecticides, and repellents" had the lowest percentage of structured drug prescriptions at only 23.4% (342/1461).

In total, 742 ATC level 5 codes (ingredients) were identified in the drug prescription data.

The structuredness of the ingredients varied widely, showing a wide range of structuredness among ingredients, as shown in Figure 5, where each of the 742 ATC level 5 codes (ingredients) is represented by a single dot. The y-axis represents the degree of the structuredness between 0% and 100%. The x-axis represents the frequency of each ATC level 5 code in data set 1, with a limitation of 85.18% (1,506,059/1,768,153) structured and evaluated unstructured data. There were only 4 ATC level 5 codes that were each used more than 45,000 times in drug prescriptions, namely N02BB02 (metamizole), B05BB01 (sodium chloride), A02BC02 (pantoprazole), and N02AA05 (oxycodone).

Table 5. Number of Anatomical Therapeutic Chemical (ATC) codes by ATC level 1 for structured, unstructured, and combined data.

ATC 1st level	Structured drug prescriptions (n=843,980), n/N (%)	Unstructured drug prescriptions for evaluated subset (n=924,173), n/N (%)	Total number (n=1,768,153), n/N (%)
N – Nervous system	197,831/322,286 (61.38)	124,455/322,286 (38.62)	322,286/1,337,565 (24.1)
B – Blood and blood forming organs	164,032/251,120 (65.32)	87,088/251,120 (34.68)	251,120/1,337,565 (18.77)
A – Alimentary tract and metabolism	137,988/250,543 (55.08)	112,555/250,543 (44.92)	250,543/1,337,565 (18.73)
C – Cardiovascular system	170,703/247,629 (68.93)	76,926/247,629 (31.07)	247,629/1,337,565 (18.51)
J – Anti-infective for systemic use	60,844/88,659 (68.63)	27,815/88,659 (31.37)	88,659/1,337,565 (6.63)
H – Systemic hormonal drugs, excluding sex hormones and insulins	51,296/64,199 (79.9)	12,903/64,199 (20.1)	64,199/1,337,565 (4.8)
M – Musculo-skeletal system	12,083/36,819 (32.82)	24,736/36,819 (67.18)	36,819/1,337,565 (2.75)
R – Respiratory system	19,686/28,148 (69.94)	8462/28,148 (30.06)	28,148/1,337,565 (2.1)
V – Various	9639/14,672 (65.7)	5033/14,672 (34.3)	14,672/1,337,565 (1.1)
L – Antineoplastic and immunomodulating agents	8670/14,538 (59.64)	5868/14,538 (40.36)	14,538/1,337,565 (1.09)
G – Genito urinary system and sex hormones	3662/8778 (41.71)	5116/8778 (58.28)	8778/1,337,565 (0.66)
S – Sensory organs	5077/5179 (98.03)	102/5179 (1.97)	5179/1,337,565 (0.39)
D – Dermatologicals	2127/3534 (60.19)	1407/3534 (39.81)	3534/1,337,565 (0.26)
P – Antiparasitic products, insecticides, and repellents	342/1461 (23.41)	1119/1461 (76.59)	1461/1,337,565 (0.11)
no med	0/1,768,153 (0)	166,307/1,768,153 (9.41)	166,307/1,768,153 (9.41)
unspec	0/1,768,153 (0)	2187/1,768,153 (0.12)	2187/1,768,153 (0.12)
Total validated	843,980/1,768,153 (47.73)	662,079/1,768,153 (37.44)	1,506,059/1,768,153 (85.18)
Not validated	0/1,768,153 (0)	262,094/1,768,153 (14.82)	262,094/1,768,153 (14.82)

Figure 4. Structuredness of drug prescriptions by ATC groups for 85.18% of initial data set DS1. ATC: Anatomical Therapeutic Chemical.

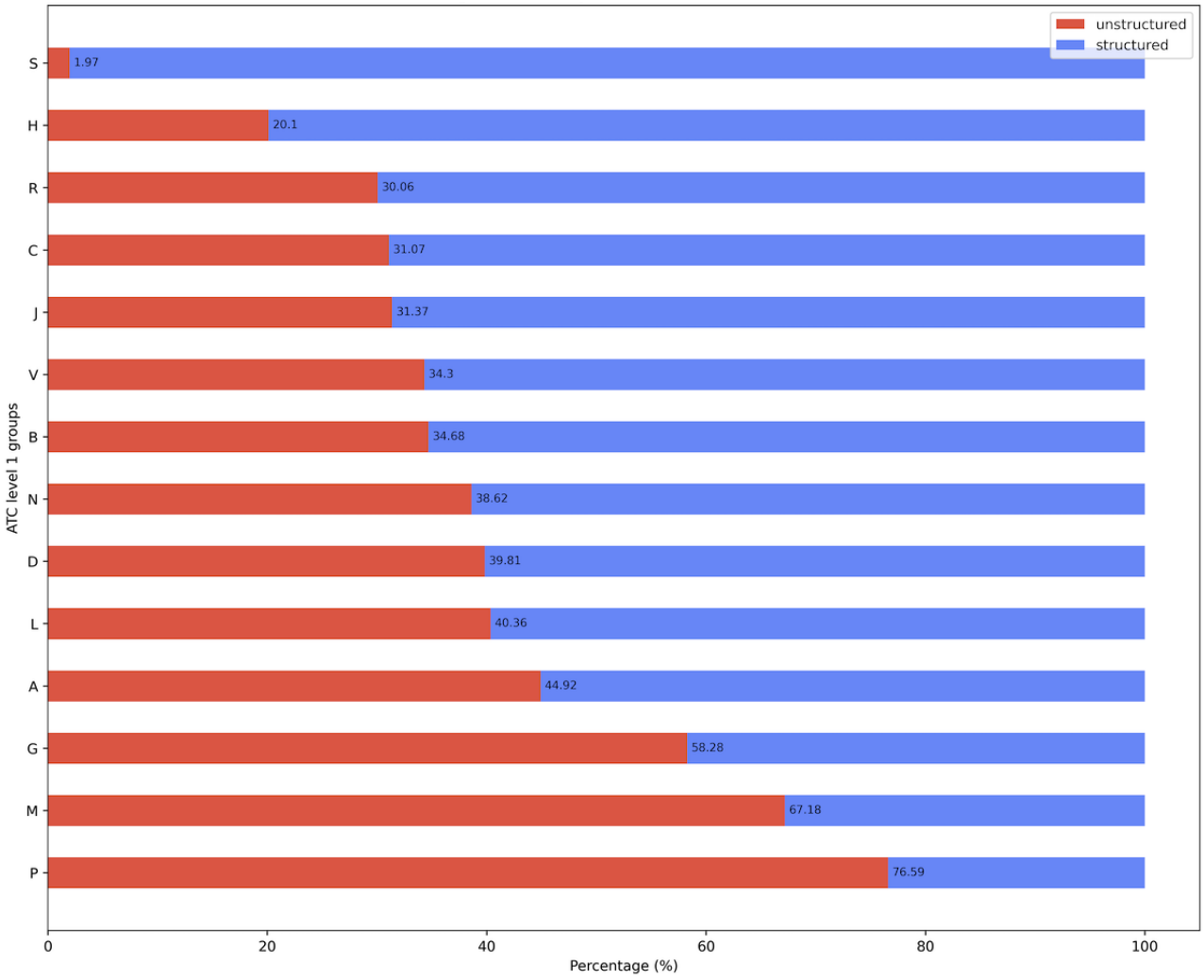
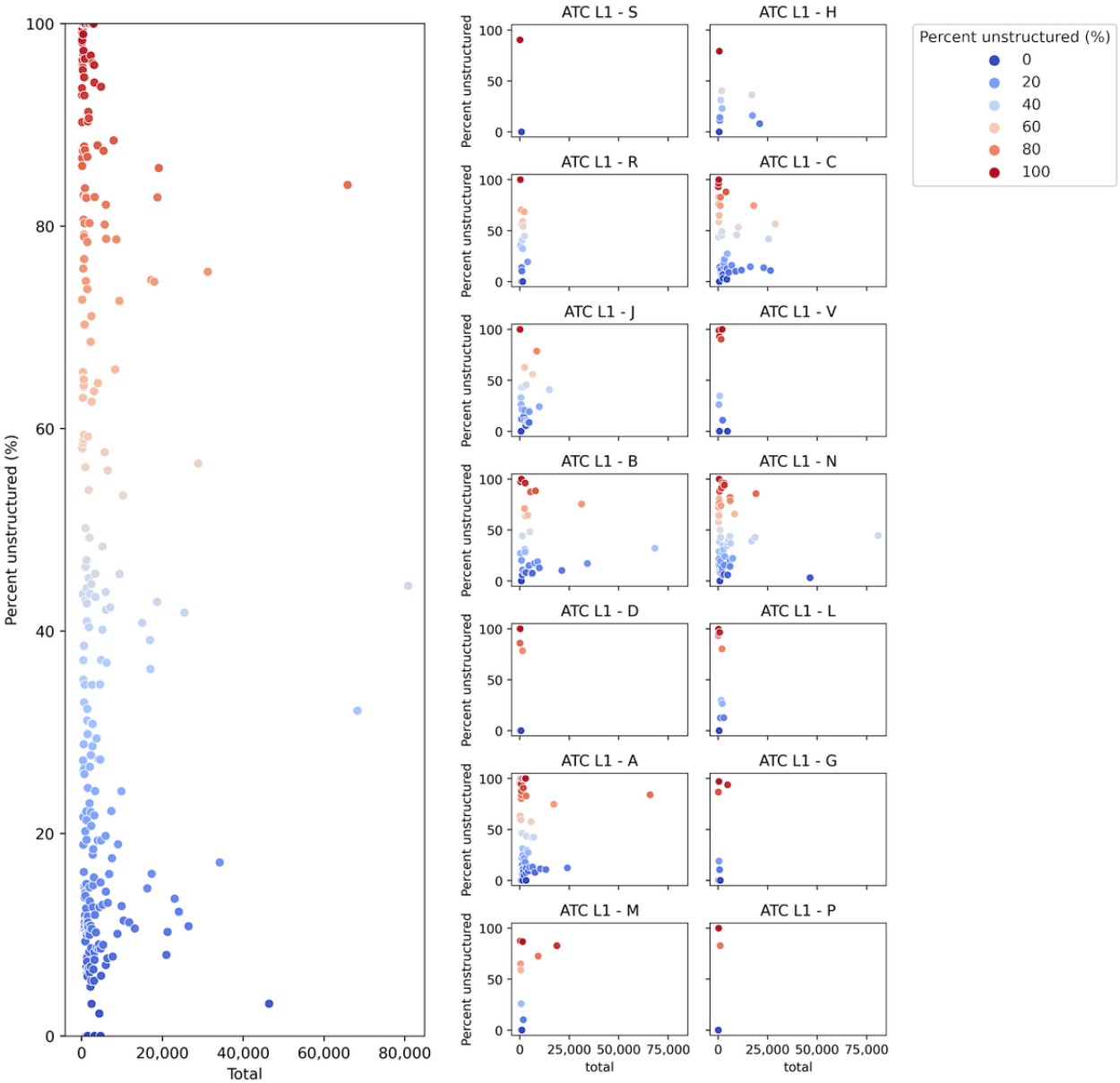


Figure 5. Structuredness of drug prescription by ATC L5 (a) total unstructured data and (b) by each ATC L1 group. A: Alimentary tract and metabolism; ATC: Anatomical Therapeutic Chemical; B: Blood and blood forming organs; C: Cardiovascular system; D: Dermatologicals; G: Genito urinary system and sex hormones; H: Systemic hormonal drugs, excluding sex hormones and insulins; J: Anti-infective for systemic use; L: Antineoplastic and immunomodulating agents; L1: level 1; L5: level 5; M: Musculo-skeletal system; N: Nervous system; P: Antiparasitic products, insecticides, and repellents; R: Respiratory system; S: Sensory organs; V: Various.



Together, these 4 ATC level 5 codes accounted for 14.79% (261,460/1,768,153) of the total data in the data set 1. Pantoprazole had the lowest level of structured data for these 4 ingredients (10,490/65,861, 15.93%), whereas oxycodone had the highest level of structured data (44,952/46,434, 96.81 %). Multimedia Appendix 3 contains the complete list of all ingredients with their ATC level 5 codes, number of structured, unstructured, and total drug prescriptions. In addition, it includes the percentage strength of each ingredient, ordered by the total number of drug prescriptions, starting with the largest number.

Discussion

Principal Findings

Our 4-step approach ensures data quality assessment as recommended by Zozus et al [18]. We provide transparency by reporting the structuredness of drug prescriptions. In addition, our approach improved the structuredness and thus the completeness of drug prescription data. This leads to better usability for secondary use on research infrastructure, such as the OMOP based ATC codes accompanied by manual review.

The initial analysis of the drug data showed a ratio of 52.3% unstructured to 47.7% structured drug prescriptions. With the algorithms presented, the structuredness could be increased to 85.1% with 1 level of evaluation. For the evaluation of the initial data set 1, manual examination of the most frequent 1000 free-text entries was sufficient, and we were able to achieve the targeted minimum coverage of 80%. Algorithm 3, which was based on similarity matching, was found to quantitatively outperform the other 2 algorithms, providing results for all unstructured drug prescriptions. In terms of the reliability of the results, algorithm 3 had a correctness rate of only 76.5%.

Therefore, the evaluation phase was critical to manually correct all incorrectly derived ATC codes. In addition, the manual evaluation process was critical in identifying patterns that can be used to determine the reliability of the algorithms based on additional factors, such as algorithm-to-algorithm matches and the Levenshtein similarity score of algorithm 3. When all 3 methods or algorithms 1 and 2 yielded the same ATC code, the algorithm results were considered correct in each case. The scoring process yielded a minuscule percentage of incorrect results (approximately 1.5%) when algorithms 2 and 3 or 1 and 3 yielded the same results.

The patterns identified are a good indication that can help increase the reliability of the results without further manual evaluation, compared with the overall correctness of 76.5% (765/1000) found for algorithm 3 results. The Levenshtein similarity score revealed another trend for algorithm 3: incorrect results had a significantly lower mean similarity score than correct results. The exceptions have been isolated to a small number of ingredients, reasons such as missing dosage form and dosage information, and cases where the same ingredients were used for single and combination drug products, resulting in separate ATC codes. The quality of the RWD has a major impact on the results of observational studies that rely on it. It is important to ensure that RWD data are suitable for use in observational studies [24] and that any limitations or quality concerns are explicitly stated [25].

Limitations

Currently, the analyzed data set is limited to inpatient drug prescriptions from the University Hospital Carl Gustav Carus Dresden. No drug prescription data from intensive care medicine were included, and no other institution has used our technique yet. Outliers or rare patterns may have gone undetected because the study was limited to the first 1000 free-text prescriptions. Although this covers most of the data, the results of the algorithm for the remaining free-text entries are yet to be evaluated. This study does not include an outcome evaluation of additional drug prescription entries based on identified patterns. Currently, the method is limited to determining ATC codes for unstructured drug prescriptions and does not consider other terminologies such as RxNorm.

Comparison With Prior Work

Most studies evaluating the quality of RWD data refer to the dimensions of completeness and accuracy compared with predefined gold standard data that vary by publication, as identified by Weiskopf et al [19]. We did not define our gold standard based on RWD sources but used the internationally recognized and widely used terminology ATC as standardized terminology and provided a method to automatically determine the appropriate ATC for drug prescription data that are unstructured and available only as free text. Wang et al [26] developed a rule-based data quality system with >6000 criteria for plausibility testing (eg, pregnancy is not plausible in male patients) but did not address data harmonization by mapping unstructured free-text data to defined terminologies for research. Unlike Schmidt et al [27] and Kahn et al [25], our study not only focus on data quality assessment but also defines the absence of structure in the data as free text without a corresponding ATC code and builds on previous research by proposing a method to improve unstructured data by automatically annotating the appropriate ATC code.

The high proportion of free-text or unstructured drug prescriptions was due to the hospital's prescription system and local conditions. According to previous studies on the data structure in RWD [28], this is a widespread challenge in Germany. However, the issues of dealing with unstructured data in EHR records that prevent interoperability are widespread as stated by Kruse et al [29,30] in their systematic reviews of existing literature on the use of EHR data that need to be addressed to ensure "fitness for use" in general. Compared with the well-established Unified Medical Language System MetaMap [31,32], which has been used by industry and academia for many years, our NLP approach focuses on a lightweight implementation. On the one hand, this limits the configuration possibilities, but on the other hand, it reduces the computational efforts and promotes the performance of the ATC code recognition. Because MetaMap focuses only on English and does not support German drug catalogs, our approach closes this gap and can be adapted to other languages as well.

Future Work

The presented 4-step approach can be applied to any RWD with unstructured data such as conditions, procedures, or test results. This approach will be tested in the future on other sites that provide drug data and product lists with ATC codes. In the next phases, further research will be conducted on pattern recognition to enable reliable prediction of the accuracy of results for specific ATC codes rather than manually checking them. In addition, new NLP-based algorithms will be implemented to improve the overall reliability of the results. Furthermore, our approach can be applied to other hospital sites that participate in the German Medical Informatics Initiative [33,34] in the following steps. Our approach is not limited to the German language. Because the only requirement is to provide a common list of ingredients or drug products for comparison with unstructured free text, this can work for any other language if compared texts are available in the same language.

Conclusions

RWD observational research requires a high level of data structuredness. Even more critical is the awareness of limitations as well as transparency of the level of structure of the data on which the research is based. Using drug prescriptions as a first use case, we were able to investigate and improve the structure of RWD, which can be applied to other RWDs in the future. Although the presented methods require manual verification to ensure that the results are correct, the methodology is promising and can be used to improve structuredness of data.

Acknowledgments

This study is part of the project MIRACUM, funded by the German Ministry of Education and Research (FKZ 01ZZ1801L).

Authors' Contributions

IR and FB worked on the conceptualization and methodology. JS and IR worked on the software. IR, SF, AF, and JS worked on the evaluation. IR, FB, JW, and JS analyzed the data. IR, FB, JS, and MW curated the data. FB and IR wrote the original draft. JW, SF, AF, and MS reviewed and edited the draft. IR and FB worked on the visualization. FB and MS were responsible for supervision. IR was responsible for project administration. All authors read and agreed to the published version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Small set of rules.
[XLSX File (Microsoft Excel File), 9 KB - medinform_v11i1e40312_app1.xlsx]

Multimedia Appendix 2
Data set 4.
[XLSX File (Microsoft Excel File), 131 KB - medinform_v11i1e40312_app2.xlsx]

Multimedia Appendix 3
Complete list of ATC L 5 with frequency and proportion of structured versus unstructured entries.
[XLSX File (Microsoft Excel File), 32 KB - medinform_v11i1e40312_app3.xlsx]

References

1. Maissenhaelter BE, Woolmore AL, Schlag PM. Real-world evidence research based on big data: motivation-challenges-success factors. Onkologe (Berl) 2018 Jun 7;24(Suppl 2):91-98 [FREE Full text] [doi: 10.1007/s00761-018-0358-3] [Medline: 30464373]
2. Brown JD. A call to action to track generic drug quality using real-world data and the FDA's sentinel initiative. J Manag Care Spec Pharm 2020 Aug;26(8):1050. [doi: 10.18553/jmcp.2020.26.8.1050] [Medline: 32715968]
3. Desai RJ, Matheny ME, Johnson K, Marsolo K, Curtis LH, Nelson JC, et al. Broadening the reach of the FDA Sentinel system: a roadmap for integrating electronic health record data in a causal analysis framework. NPJ Digit Med 2021 Dec 20;4(1):170 [FREE Full text] [doi: 10.1038/s41746-021-00542-0] [Medline: 34931012]
4. Cocoros NM, Fuller CC, Adimadhyam S, Ball R, Brown JS, Dal Pan GJ, And the FDA-Sentinel COVID-19 Working Group. A COVID-19-ready public health surveillance system: The Food and Drug Administration's Sentinel System. Pharmacoepidemiol Drug Saf 2021 Jul 18;30(7):827-837 [FREE Full text] [doi: 10.1002/pds.5240] [Medline: 33797815]
5. Behrman RE, Benner JS, Brown JS, McClellan M, Woodcock J, Platt R. Developing the Sentinel System--a national resource for evidence development. N Engl J Med 2011 Feb 10;364(6):498-499. [doi: 10.1056/NEJMp1014427] [Medline: 21226658]
6. Robb MA, Racoosin JA, Sherman RE, Gross TP, Ball R, Reichman ME, et al. The US Food and Drug Administration's Sentinel Initiative: expanding the horizons of medical product safety. Pharmacoepidemiol Drug Saf 2012 Jan 19;21 Suppl 1:9-11. [doi: 10.1002/pds.2311] [Medline: 22262587]
7. Ball R, Robb M, Anderson SA, Dal Pan G. The FDA's sentinel initiative--A comprehensive approach to medical product surveillance. Clin Pharmacol Ther 2016 Mar;99(3):265-268. [doi: 10.1002/cpt.320] [Medline: 26667601]
8. Platt R, Brown JS, Robb M, McClellan M, Ball R, Nguyen MD, et al. The FDA sentinel initiative — an evolving national resource. N Engl J Med 2018 Nov 29;379(22):2091-2093. [doi: 10.1056/nejmp1809643]
9. Gini R, Sturkenboom MC, Sultana J, Cave A, Landi A, Pacurariu A, Working Group 3 of ENCePP (Inventory of EU data sourcesmethodological approaches for multisource studies). Different strategies to execute multi-database studies for medicines surveillance in real-world setting: a reflection on the European model. Clin Pharmacol Ther 2020 Aug;108(2):228-235 [FREE Full text] [doi: 10.1002/cpt.1833] [Medline: 32243569]
10. EHDEN Homepage. European Health Data & Evidence Network. URL: https://www.ehden.eu/ [accessed 2022-05-05]
11. COVID-19: EMA sets up infrastructure for real-world monitoring of treatments and vaccines. European Medicines Agency. 2020 Jul 21. URL: https://www.ema.europa.eu/en/news/covid-19-ema-sets-infrastructure-real-world-monitoring-treatments-vaccines [accessed 2022-01-13]
12. Hripcsak G, Schuemie MJ, Madigan D, Ryan PB, Suchard MA. Drawing reproducible conclusions from observational clinical data with OHDSI. Yearb Med Inform 2021 Aug;30(1):283-289 [FREE Full text] [doi: 10.1055/s-0041-1726481] [Medline: 33882595]

13. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. Stud Health Technol Inform 2015;216:574-578 [FREE Full text] [Medline: 26262116]
14. Garza M, Del Fiol G, Tenenbaum J, Walden A, Zozus MN. Evaluating common data models for use with a longitudinal community registry. J Biomed Inform 2016 Dec;64:333-341 [FREE Full text] [doi: 10.1016/j.jbi.2016.10.016] [Medline: 27989817]
15. Reinecke I, Zoch M, Reich C, Sedlmayr M, Bathelt F. The usage of OHDSI OMOP - A scoping review. Stud Health Technol Inform 2021 Sep 21;283:95-103. [doi: 10.3233/SHTI210546] [Medline: 34545824]
16. Introduction to Drug Utilization Research. Geneva: World Health Organization; 2003.
17. Anatomical Therapeutic Chemical (ATC) Classification. World Health Organization. URL: https://www.who.int/tools/atc-ddd-toolkit/atc-classification [accessed 2022-05-05]
18. Zozus MW, Hammond E, Green GG, Kahn MG, Richesson RL, Rusincovitch SA, et al. Data quality assessment recommendations for secondary ise of EHR data. Research Gate. 2015 Oct. URL: https://www.researchgate.net/publication/283267713_Data_Quality_Assessment_Recommendations_for_Secondary_ise_of_EHR_Data [accessed 2022-05-05]
19. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc 2013 Jan 01;20(1):144-151 [FREE Full text] [doi: 10.1136/amiajnl-2011-000681] [Medline: 22733976]
20. Fuzzy String Matching in Python. GitHub. URL: https://github.com/seatgeek/fuzzywuzzy [accessed 2021-11-26]
21. Hutchison E, Zhang Y, Nampally S, Weatherall J, Khan F, Shameer K. Uncovering machine learning-ready data from public clinical trial resources: a case-study on normalization across aggregate content of ClinicalTrials.Gov24322. In: Proceedings of the 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 2020 Presented at: 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); Dec 16-19, 2020; Seoul, Korea (South). [doi: 10.1109/bibm49941.2020.9313362]
22. Bobroske K, Larish C, Cattrell A, Bjarnadóttir MV, Huan L. The bird's-eye view: a data-driven approach to understanding patient journeys from claims data. J Am Med Inform Assoc 2020 Jul 01;27(7):1037-1045 [FREE Full text] [doi: 10.1093/jamia/ocaa052] [Medline: 32521006]
23. Reinecke I. source code of the implemented algorithm and visualization. ResearchGate. 2022. URL: https://www.researchgate.net/publication/366867005_drug_data-publication [accessed 2023-01-04]
24. Defeo JA. Juran's Quality Handbook: The Complete Guide to Performance Excellence, Seventh Edition. New York: McGraw Hill Education; Nov 11, 2016.
25. Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, et al. A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. EGEMS (Wash DC) 2016 Sep 11;4(1):1244 [FREE Full text] [doi: 10.13063/2327-9214.1244] [Medline: 27713905]
26. Wang Z, Talburt JR, Wu N, Dagtas S, Zozus MN. A rule-based data quality assessment system for electronic health record data. Appl Clin Inform 2020 Aug;11(4):622-634 [FREE Full text] [doi: 10.1055/s-0040-1715567] [Medline: 32968999]
27. Schmidt CO, Struckmann S, Enzenbach C, Reineke A, Stausberg J, Damerow S, et al. Facilitating harmonized data quality assessments. A data quality framework for observational health research data collections with software implementations in R. BMC Med Res Methodol 2021 Apr 02;21(1):63 [FREE Full text] [doi: 10.1186/s12874-021-01252-7] [Medline: 33810787]
28. Vass A, Reinecke I, Boeker M, Prokosch H, Gulden C. Availability of structured data elements in electronic health records for supporting patient recruitment in clinical trials. In: Studies in Health Technology and Informatics. Amsterdam, Netherlands: IOS Press; 2021.
29. Kruse CS, Kristof C, Jones B, Mitchell E, Martinez A. Barriers to electronic health record adoption: a systematic literature review. J Med Syst 2016 Dec;40(12):252 [FREE Full text] [doi: 10.1007/s10916-016-0628-9] [Medline: 27714560]
30. Kruse CS, Stein A, Thomas H, Kaur H. The use of electronic health records to support population health: a systematic review of the literature. J Med Syst 2018 Sep 29;42(11):214 [FREE Full text] [doi: 10.1007/s10916-018-1075-6] [Medline: 30269237]
31. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp 2001:17-21 [FREE Full text] [Medline: 11825149]
32. Aronson AR, Lang F. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc 2010;17(3):229-236 [FREE Full text] [doi: 10.1136/jamia.2009.002733] [Medline: 20442139]
33. Gehring S, Eulenfeld R. German medical informatics initiative: unlocking data for research and health care. Methods Inf Med 2018 Jul 17;57(S 01):e46-e49. [doi: 10.3414/me18-13-0001]
34. Semler S, Wissing F, Heyder R. German medical informatics initiative. Methods Inf Med 2018 Jul 17;57(S 01):e50-e56. [doi: 10.3414/me18-03-0003]

Abbreviations

ATC: Anatomical Therapeutic Chemical

EHR: electronical health record
NLP: natural language processing
OMOP: Observational Medical Outcomes Partnership
RCT: randomized controlled trial
RWD: real-world data
UKD: University Hospital Carl Gustav Carus Dresden

Edited by C Lovis; submitted 15.06.22; peer-reviewed by A Lamer, M Pedrera Jiménez, B Ru; comments to author 07.09.22; revised version received 27.09.22; accepted 18.11.22; published 25.01.23.

Please cite as:
Reinecke I, Siebel J, Fuhrmann S, Fischer A, Sedlmayr M, Weidner J, Bathelt F
Assessment and Improvement of Drug Data Structuredness From Electronic Health Records: Algorithm Development and Validation
JMIR Med Inform 2023;11:e40312
URL: https://medinform.jmir.org/2023/1/e40312
doi:10.2196/40312
PMID:36696159

Original Paper

An Ontology-Based Approach for Consolidating Patient Data Standardized With European Norm/International Organization for Standardization 13606 (EN/ISO 13606) Into Joint Observational Medical Outcomes Partnership (OMOP) Repositories: Description of a Methodology

Santiago Frid[1,2], MSc, MD; Xavier Pastor Duran[1,2], MD, PhD; Guillem Bracons Cucó[3], MSc; Miguel Pedrera-Jiménez[4], MSc, PhD; Pablo Serrano-Balazote[5], MD; Adolfo Muñoz Carrero[6], PhD; Raimundo Lozano-Rubí[1,2], MD, PhD

[1]Medical Informatics Unit, Hospital Clínic de Barcelona, Barcelona, Spain

[2]Clinical Foundations Department, Universitat de Barcelona, Barcelona, Spain

[3]Fundació Clínic per a la Recerca Biomédica, Barcelona, Spain

[4]Data Science Unit, Hospital 12 de Octubre, Madrid, Spain

[5]Direction of Planification, Hospital 12 de Octubre, Madrid, Spain

[6]Unit of Investigation in Telemedicine and Digital Health, Instituto de Salud Carlos III, Madrid, Spain

Corresponding Author:
Santiago Frid, MSc, MD
Clinical Foundations Department
Universitat de Barcelona
Casanova 143
Barcelona, 08036
Spain
Phone: 34 934035258
Email: santifrid@gmail.com

Abstract

Background: To discover new knowledge from data, they must be correct and in a consistent format. OntoCR, a clinical repository developed at Hospital Clínic de Barcelona, uses ontologies to represent clinical knowledge and map locally defined variables to health information standards and common data models.

Objective: The aim of the study is to design and implement a scalable methodology based on the dual-model paradigm and the use of ontologies to consolidate clinical data from different organizations in a standardized repository for research purposes without loss of meaning.

Methods: First, the relevant clinical variables are defined, and the corresponding European Norm/International Organization for Standardization (EN/ISO) 13606 archetypes are created. Data sources are then identified, and an extract, transform, and load process is carried out. Once the final data set is obtained, the data are transformed to create EN/ISO 13606–normalized electronic health record (EHR) extracts. Afterward, ontologies that represent archetyped concepts and map them to EN/ISO 13606 and Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) standards are created and uploaded to OntoCR. Data stored in the extracts are inserted into its corresponding place in the ontology, thus obtaining instantiated patient data in the ontology-based repository. Finally, data can be extracted via SPARQL queries as OMOP CDM–compliant tables.

Results: Using this methodology, EN/ISO 13606–standardized archetypes that allow for the reuse of clinical information were created, and the knowledge representation of our clinical repository by modeling and mapping ontologies was extended. Furthermore, EN/ISO 13606–compliant EHR extracts of patients (6803), episodes (13,938), diagnosis (190,878), administered medication (222,225), cumulative drug dose (222,225), prescribed medication (351,247), movements between units (47,817), clinical observations (6,736,745), laboratory observations (3,392,873), limitation of life-sustaining treatment (1,298), and procedures (19,861) were created. Since the creation of the application that inserts data from extracts into the ontologies is not yet finished, the queries were tested and the methodology was validated by importing data from a random subset of patients into the ontologies using a locally developed Protégé plugin (“OntoLoad”). In total, 10 OMOP CDM–compliant tables (“Condition_occurrence,”

864 records; “Death,” 110; “Device_exposure,” 56; “Drug_exposure,” 5609; “Measurement,” 2091; “Observation,” 195; “Observation_period,” 897; “Person,” 922; “Visit_detail,” 772; and “Visit_occurrence,” 971) were successfully created and populated.

Conclusions: This study proposes a methodology for standardizing clinical data, thus allowing its reuse without any changes in the meaning of the modeled concepts. Although this paper focuses on health research, our methodology suggests that the data be initially standardized per EN/ISO 13606 to obtain EHR extracts with a high level of granularity that can be used for any purpose. Ontologies constitute a valuable approach for knowledge representation and standardization of health information in a standard-agnostic manner. With the proposed methodology, institutions can go from local raw data to standardized, semantically interoperable EN/ISO 13606 and OMOP repositories.

(JMIR Med Inform 2023;11:e44547) doi:10.2196/44547

KEYWORDS

health information interoperability; health research; health information standards; dual model; secondary use of health data; Observational Medical Outcomes Partnership Common Data Model; European Norm/International Organization for Standardization 13606; health records; ontologies; clinical data

Introduction

The term primary use of health data encompasses the generation and use of data within the context of individual health care in hospitals and physicians’ offices to serve direct care needs [1]. The term secondary use of health data is defined by the American Medical Informatics Association as “non-direct care use of PHI [personal health information] including but not limited to analysis, research, quality/safety measurement, public health, payment, provider certification or accreditation, and marketing and other business including strictly commercial activities” [2]. Although they can be further categorized [3], one of the main types of secondary uses is research.

Clinical data sharing for research is highly relevant from a scientific, economic, and ethical perspective [4]. The overwhelming increment in the volume of available data is directly related with the emergence of a new paradigm of scientific methodology in which massive amounts of data are processed and analyzed for obtaining knowledge through machine learning and data mining algorithms [5].

Despite the growth of big data technologies and the use of artificial intelligence, in order to discover new knowledge from data, they must be correct and in a consistent format, which requires a great amount of resources for cleaning, binding, and organizing them. The semantics of data is a key component regarding the aforementioned challenges. To use the electronic health record (EHR) data for different projects, it must maintain its semantics and context, independently of any particular use case. This is especially important in research, where EHR reuse processes are often based on black boxes on which the final data customer is unaware of how the data uploaded to their research database were recorded, extracted, and transformed [6].

A common health information standard should be used in both primary and secondary use to share clinical information in a way that it can be unequivocally interpreted, both syntactically and semantically, by 2 or more systems. European Norm/International Organization for Standardization (EN/ISO) 13606 is a health information standard that seeks to define a rigorous and stable architecture for communicating the health records of a single patient, preserving the original clinical meaning. It is based on a dual model that includes a reference model (RM; with the necessary components and their constraints to represent EHR extracts) and an archetype model (AM; for the formalization of clinical-domain concepts according to the RM) [7,8]. Archetypes allow the formal representation of the structure of clinical information and its meaning (through terminology binding) so that it is automatically processable by information systems.

Furthermore, the EN/ISO 13940 norm [9] provides a conceptual framework centered in the clinical process. This norm, based on a clinical perspective, defines the system of concepts that are necessary for achieving continuity in the caregiving process, including both the content and the context of the health activities. This ample norm defines the concepts relative to health care actors, health problems, sanitary activities, health care processes, sanitary planification, time-related concepts, responsibilities, and information management.

Moreover, the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) defines a common format (data model), as well as a common representation (terminologies, vocabularies, and coding schemes), to allow systematic analyses of disparate observational databases using a library of standard analytic routines that have been written based on the common format [10]. The OMOP CDM is considered by several authors as the most adequate data model for sharing data in EHR-based longitudinal studies [11-13].

This paper describes the work carried out between Hospital Clínic de Barcelona (HCB), Hospital 12 de Octubre (H12O), and Instituto de Salud Carlos III (ISCIII), which seeks to consolidate clinical data of hospitalized patients with COVID-19 from different hospitals in joint repositories, structured with EN/ISO 13606 and then normalized according to the OMOP CDM.

The aim of this study is to design and implement a scalable methodology based on the dual-model paradigm and the use of ontologies to consolidate clinical data from different organizations in a standardized repository for research purposes without loss of meaning. This implies a series of particular objectives such as (1) to define a set of relevant clinical and

biochemical variables of patients hospitalized with COVID-19, (2) to model a set of standardized archetypes based on EN/ISO 13606 to communicate such information, (3) to conceptually represent those clinical variables by means of ontologies, (4) to generate EN/ISO 13606–standardized EHR extracts of COVID-19 patients, and (5) to map and transform the source data to create OMOP CDM–compliant tables.

Methods

Ethical Considerations

This study was approved by the Hospital Clínic de Barcelona Ethics Committee for Investigation with Drugs (HCB/2018/0573).

Cohort Inclusion Criteria

We included in this project patients with COVID-19 admitted to the emergency room (ER) or hospitalized between February 17, 2020 (beginning of the first wave in Spain), and February 15, 2022 (end of the sixth wave in Spain).

Methodology

The following methodology comprises a series of steps in order to achieve the study's objectives.

Step 1: Definition of Clinical Variables, Data Structures, and EN/ISO 13606 Archetypes

The first step consists in deciding the clinically relevant variables that should be included. Afterward, the data structures must be defined, including the fields, their descriptions, and the standardized terminologies and classifications to be used.

Since OMOP CDM is intended for secondary use of data (specifically, for biomedical research), its granularity is somewhat reduced when compared to raw data captured in hospitals. For this reason, the Medical Informatics Unit (MIU) at HCB decided to first standardize the data according to EN/ISO 13606, in order to have semantically interoperable EHR extracts with the maximum level of detail.

Therefore, the MIU at HCB and the Data Science Unit at H12O defined the EN/ISO 13606 archetypes to be used, modeled with the software LinkEHR [14] created by VeraTech for Health. The data types used are those established by the standard's RM.

This RM has multiple components, including the entry (a result of 1 clinical action, 1 observation, 1 clinical interpretation, or 1 intention) and its elements (the leaf node of the EHR hierarchy, containing a single data value). In our project, the archetypes modeled at the entry level of the RM were the following: diagnosis, episodes, limitation of life-sustaining treatment, administered medication, cumulative drug dose, prescribed medication, movements between units, clinical observations, laboratory observations, patients, health problems, and procedures. These archetypes were registered under a Creative Commons license (ID 2204210968527), so that any user who follows the license terms can share and adapt them [15]. Figure 1 shows a mind map of the diagnosis entry archetype as an example.

Figure 1. Mind map of the EN/ISO 13606 "diagnosis" archetype in Spanish, modeled with LinkEHR. The "diagnosis" entry has 6 elements: episode_id, diagnosis, diagnosis_datetime, patient_id, diagnosis_id, and source. Each of them has its corresponding data type. EN/ISO: European Norm/International Organization for Standardization.

Step 2: Identification of Data Sources and Extract, Transform, and Load

Afterward, the corresponding data sources must be identified, in order to carry out the extract, transform, and load (ETL) process. In our case, these sources were (1) structured data from HCB's health information system (HIS), SAP; (2) unstructured data from HCB's HIS. A collaborative work with Barcelona Supercomputing Center (BSC) allows for the recognition of clinical entities through natural language processing and its extraction as normalized structured data; (3) outpatient setting structured data from Agència de Qualitat i Avaluació Sanitàries de Catalunya.

Since the last 2 sources come from separate projects whose description is besides the objective of this paper, we will focus on the first one. Archetypes created in the previous step were used as templates for identifying data in the aforementioned sources. Periodic meetings were held with the Information Technology Department at HCB to identify the location of the data and the transformations needed to obtain the structured data defined in the previous step. Once this was achieved, the tables were loaded into a MySQL database hosted on a dedicated server of the MIU.

Step 3: Creation of EN/ISO 13606 EHR Extracts From Source Data

Once the final data set is obtained, data must be transformed to create EHR extracts normalized according to EN/ISO 13606. This transformation includes mapping of local variables to standardized nomenclatures and classifications (Systematized Nomenclature of Medicine—Clinical Terms (SNOMED CT), International Classification of Diseases 10—Clinical Modification (ICD-10-CM), Logical Observation Identifiers Names and Codes (LOINC)), assigning readable descriptions to local codes, and categorizing certain concepts (eg, grouping hospital units according to the level of care).

This process is performed by mapping archetypes to the implicated information systems, without the need to modify them. This approach allows the automation of data extraction and the reuse of this methodology for other use cases with very little effort, which constitutes one of the great advantages of dual-model strategies.

In our case, we carried out this process with the help of VeraTech for Health, our technical partners, using LinkEHR, thus creating extracts on our dedicated server and constituting an EN/ISO 13606 standardized clinical repository. Figure 2 shows a test example of an EN/ISO 13606 EHR extract (without real-patient data). In this extract, the ICD-10-CM code H40.9 (unspecified glaucoma) is being communicated, alongside its date and time of record and the ID of the clinical episode it pertains to.

Figure 2. Anonymized, normalized EN/ISO 13606 EHR extract of diagnosis in Spanish. EHR: electronic health record; EN/ISO: European Norm/International Organization for Standardization.

```
<content xsi:type="SECTION">
    <rc_id extension="842EB49C-3C3E-4ED3-80D3-89FDEB560004" root="clinic" xsi:type="INSTANCE_IDENTIFIER"/>
    <archetype_id extension="at0015" identifier_name="DIAGNOSTICOS" root="ISO-EN13606-EHR_EXTRACT.ExtractoCOVID19.v1" xsi:type="INSTANCE_IDENTIFIER"/>
    <members xsi:type="ENTRY">
        <rc_id extension="A54250CE-B2B8-4337-913C-0690CAC87E4F" root="clinic" xsi:type="INSTANCE_IDENTIFIER"/>
        <archetype_id extension="at0016" identifier_name="Diagnostico" root="ISO-EN13606-EHR_EXTRACT.ExtractoCOVID19.v1" xsi:type="INSTANCE_IDENTIFIER"/>
        <items xsi:type="ELEMENT">
            <rc_id extension="1E6A2DA3-10CD-4BD6-83BD-26A4D632A571" root="clinic" xsi:type="INSTANCE_IDENTIFIER"/>
            <archetype_id extension="at0017" identifier_name="episodio_id" root="clinic" xsi:type="INSTANCE_IDENTIFIER"/>
            <value extension="9995689660" root="clinic" xsi:type="INSTANCE_IDENTIFIER"/>
        </items>
        <items xsi:type="ELEMENT">
            <rc_id extension="857FEC7D-83EB-41D9-9C2B-DCA732F7B800" root="clinic" xsi:type="INSTANCE_IDENTIFIER"/>
            <archetype_id extension="at0018" identifier_name="diagnostico" root="clinic" xsi:type="INSTANCE_IDENTIFIER"/>
            <value code="H40.9" code_system="ICD10" xsi:type="CODED_VALUE"/>
        </items>
        <items xsi:type="ELEMENT">
            <rc_id extension="9881D930-540A-46D2-AF08-C98D17B372DA" root="clinic" xsi:type="INSTANCE_IDENTIFIER"/>
            <archetype_id extension="at0019" identifier_name="fecha_hora_Dx" root="clinic" xsi:type="INSTANCE_IDENTIFIER"/>
            <value xsi:type="DATE_TIME" value="2019-03-04T10:01:38"/>
        </items>
```

Step 4: Creation of Ontologies

Traditionally, clinical concepts and the relationships between them have been poorly developed in HISs. The MIU at HCB developed OntoCR, an ontology-based clinical repository, conforming to EN/ISO 13606 standard [16,17]. The use of ontologies allows for the definition of a conceptual architecture centered on the representation of the clinical process, while the use of EN/ISO 13606 allows syntactic and semantic interoperability between systems. EN/ISO 13940 was also used to define the generic concepts needed to achieve continuity of care, representing both the content and the context of the health care services.

One of the main advantages of ontologies is their flexibility to perform changes with minimum use of resources, adapting to an ever-changing environment. Likewise, ontologies allow the addition of conceptual layers, thus mapping locally defined concepts to health information standards, facilitating the communication of information without loss of meaning.

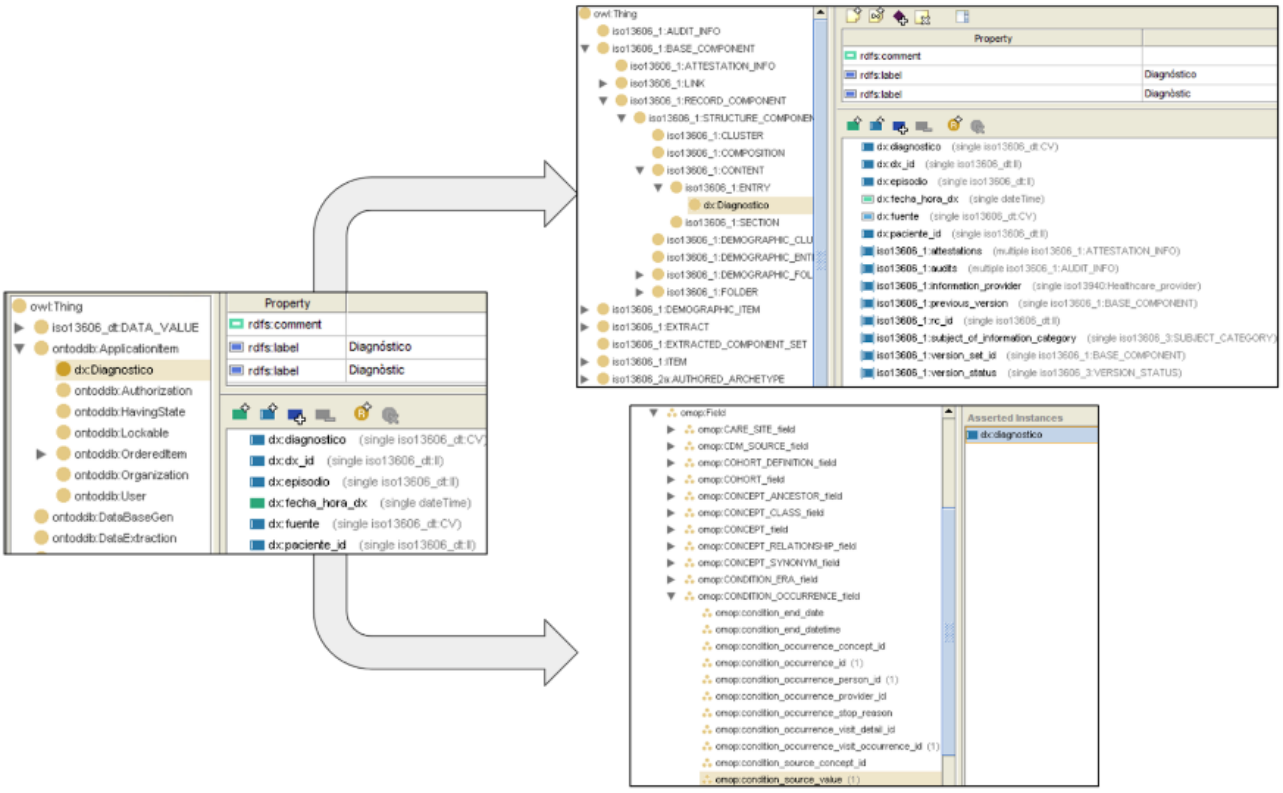
A relational database (OWL-DB) is used for storing ontologies and instantiated data, designed according to the Web Ontology Language (OWL) specification [18]. The ontologies in this project were created using Protégé, a free, open-source ontology editor created by Stanford University that fully supports OWL and Resource Description Framework (RDF) specifications from the World Wide Web Consortium [19]. A plug-in developed by our team, the OWL-DB plugin, connects Protégé with the OWL-DB module at the storage level.

These ontologies were conceptualized in 3 different layers. The first one describes the concepts modeled in the archetypes, with the classes and properties that describe the data structure defined in the first phase. Data types according to the EN/ISO 13606 RM were used.

In the next layer, we used a locally created ontology that reproduces the EN/ISO 13606 RM and AM. By creating an additional ontology that maps the archetypal concepts to the EN/ISO 13606 model, we structured our data according to the standard. In this layer, each entry-level archetype is represented in a separate ontology.

As with EN/ISO 13606, we created an ontology that models the OMOP CDM and afterward mapped archetypal concepts to the corresponding meta-class of the standard. So, the third layer consists of ontologies for each archetype that reproduce concepts according to the OMOP CDM structure. Figure 3 shows these 3 ontologies. The left image (ontology of the AM of diagnosis) depicts the class "Diagnosis" with its properties diagnosis, diagnosis_id, episode, diagnosis_datetime, source, and patient_id. In the upper-right image, a new ontology was created where the class "Diagnosis" was modeled as a subclass of "iso13606: Entry," thus inheriting its properties defined in the RM. Finally, in the lower right image, a third ontology maps the property diagnosis with OMOP CDM's metaclass "condition_source_value."

Figure 3. Ontologies of the archetype model of diagnosis (left) and its mapping to the EN/ISO 13606 structure (upper right) and the OMOP CDM (lower right) in Spanish, edited with Protégé software. EN/ISO: European Norm/International Organization for Standardization; OMOP CDM: Observational Medical Outcomes Partnership Common Data Model.



Afterward, these ontologies must be loaded into a production environment of OntoCR so as to generate the structure that can receive instantiated data of patients and store it.

Step 5: Integration of EN/ISO 13606 Extracts Into the Ontology-Based Clinical Repository and Extraction of Data as OMOP CDM–Compliant Tables

Once the ontological structure is ready to receive the data, the EHR extracts must be inserted into the repository, thus incorporating the normalized, instantiated data. We initially explored the possibility of adapting a preexisting application programming interface (API) that was used for the same purpose in a previous project. However, the resources needed for its adaptation were significantly elevated, and its scalability reduced. Therefore, we decided to work on an application that identifies each archetype node within the extract and inserts it into its counterpart in the OWL file. This is facilitated by the representation in the ontologies of each archetype, their nodes, and the data types used (compliant with EN/ISO 13606).

Finally, data stored in the ontology-based clinical repository needs to be extracted through SPARQL queries, a language used for graph databases. Since archetypal concepts have been previously mapped to the OMOP CDM, by performing these queries, the extraction process is simplified. If there are cases in which data needs to be transformed to fit the CDM, such transformations can be included in the queries or carried out via SQL queries once relational tables are obtained.

In Figure 4, a SPARQL query for extracting data for the OMOP CDM PERSON table is shown. Attributes that are not present in the ontological repository must still be included in the SELECT clause so as to create the corresponding table column without any instantiated data. Since EN/ISO 13606 data types were used in the extracts and modeled in the ontologies, they were also represented in the queries (see the lower lines of SPARQL code).

Figure 4. SPARQL query for the "Person" table of the OMOP CDM. OMOP CDM: Observational Medical Outcomes Partnership Common Data Model.

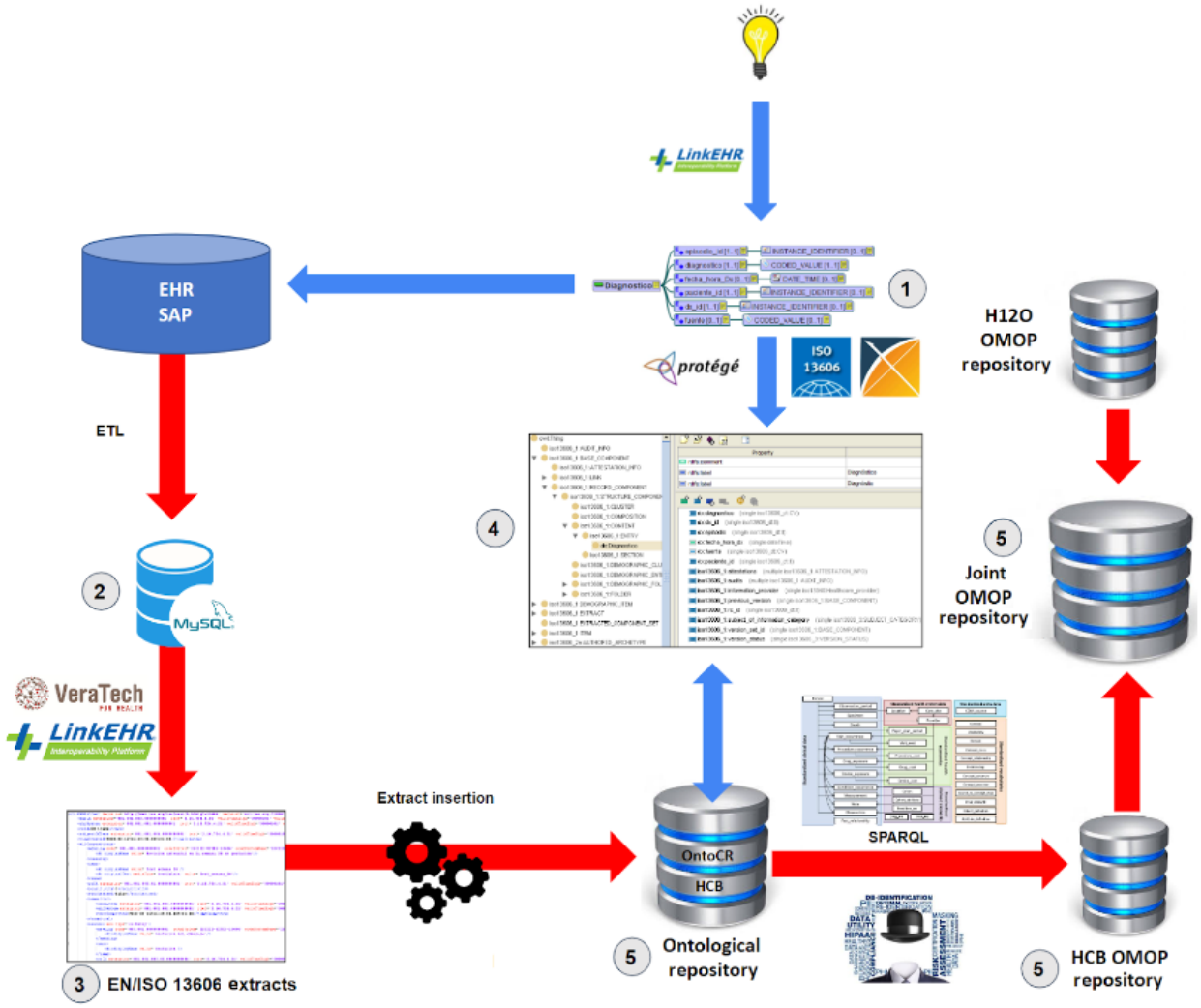
```
PREFIX odd: <http://ontoar.clinic.cat/ontologias/OMOP-CDM.Paciente.v1.owl#>
PREFIX pacn: <http://ontoar.clinic.cat/ontologias/paciente.owl#>
PREFIX omop: <http://ontoks.clinic.cat/ontologias/omop.owl#>
PREFIX iso13606_dt: <http://ontoar.clinic.cat/ontologias/iso_13606_dt.owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT ?person_id ?gender_concept_id ?year_of_birth ?month_of_birth
?day_of_birth ?birth_datetime ?race_concept_id ?ethnicity_concept_id
?location_id ?provider_id ?care_site_id ?person_source_value
?gender_source_value ?gender_source_concept_id ?race_source_value
?race_source_concept_id ?ethnicity_source_value ?ethnicity_source_concept_id
WHERE {
    ?Person_subclass rdfs:subClassOf* omop:PERSON
    OPTIONAL {?P1 rdf:type omop:person_id .}
    OPTIONAL {?P2 rdf:type omop:gender_concept_id .}
    OPTIONAL {?P3 rdf:type omop:birth_datetime .}
    OPTIONAL {?P4 rdf:type omop:gender_source_value .}
    ?Person rdf:type ?Person_subclass;
    OPTIONAL{ ?Person ?P1 ?P1_person_id .}
    OPTIONAL{ ?Person ?P2 ?P2_gender_concept_id .}
    OPTIONAL{ ?Person ?P3 ?P3_birth_datetime .}
    OPTIONAL{ ?Person ?P4 ?P4_gender_source_value .}
    ?P1_person_id iso13606_dt:identifier_name ?person_id.
    ?P2_gender_concept_id iso13606_dt:identifier_name ?gender_source_value.
    ?P4_gender_source_value iso13606_dt:code ?gender_source_value.
```

Data anonymization is performed by the IT department at this level using an institutional software solution. This way, EN/ISO 13606 extracts contain identified data that can be used for primary uses, while OMOP CDM tables are anonymized for secondary uses.

Once obtained, the anonymized data can be consolidated in a joint OMOP CDM repository with other institutions that use the same standard (in our case, H12O). OMOP CDM has a large number of tables, divided into 6 groups: standardized clinical data, standardized health system data, standardized derived elements, standardized health economics, standardized metadata, and standardized vocabularies. Our OMOP CDM repository contains the following tables, which are part of the standardized clinical data: "Condition_occurrence," "Death," "Device_exposure," "Drug_exposure," "Measurement," "Observation," "Observation_period," "Person," "Visit_detail," and "Visit_occurrence."

Figure 5 shows an overview of the whole process. The knowledge modeling starts with the creation of EN/ISO 13606 archetypes based on clinical concepts, which are then represented in ontologies that map them to EN/ISO 13606 RM and OMOP CDM. These ontologies are uploaded to OntoCR without instantiated patient data yet.

Figure 5. General overview of the process. Red arrows indicate data flow, while blue arrows indicate knowledge-related processes. Numbers indicate the deliverables within each step. EHR: electronic health record; EN/ISO: European Norm/International Organization for Standardization; ETL: extract, transform, and load; H12O: Hospital 12 de Octubre; HCB: Hospital Clínic de Barcelona; OMOP: Observational Medical Outcomes Partnership.



The data-related processes begin with the archetype-based extraction of raw data from our local system into a MySQL database and its transformation to create EN/ISO EHR extracts, which are then inserted into OntoCR via an application specifically developed for this project. SPARQL queries are performed against this ontological repository to obtain an OMOP CDM repository that is consolidated with H12O in a joint one.

Results

Methodology

The main deliverable of this project is the methodology described in the previous section. By following the aforementioned steps, any health care institution can go from local raw data to standardized, semantically interoperable EN/ISO 13606 and OMOP repositories. This methodology also led to the creation of 12 EN/ISO 13606-standardized archetypes that model important clinical variables in the ER and hospitalization settings, allowing the reuse of clinical information by using it in accordance with the Creative Commons terms.

Ontologies

Another interesting result of this study is the development of the ontologies that represent OMOP CDM, as well as their mappings to EN/ISO 13606 AM and RM. This process was carried out by members of the MIU at HCB after carefully reading the pertinent documentation of these standards and designing the optimal way of using them to represent clinical concepts.

Furthermore, representing clinical variables by means of ontologies is another way of reusing clinical information. With the creation of new ontologies for each project at HCB, where we have developed the ontology-based clinical repository OntoCR, we continue to extend our clinical knowledge representation.

EN/ISO 13606 Extracts

Table 1 shows the correspondence between EHR archetypes, OMOP CDM tables, number of extracts created throughout the study, and the number of COVID-19 patients they pertain to. We have included the diagnoses recorded in the episodes of the study period as well as the historical ones. Health problem

entries are part of the aforementioned project with BSC to extract clinical entities from unstructured texts through natural language processing, so they will not be included in this table.

Table 1. Correspondence between EHR[a] archetypes, OMOP CDM[b] tables, number of extracts created throughout the study, and the number of patients they pertain to.

EHR archetype	OMOP CDM table	EN/ISO[c] 13606 extracts, n	Patients, n
Patient	"Person"	6803	6803
Episode	"Visit_occurence"	13,938	6791
Diagnosis	"Condition_occurrence"	190,878	6799
Cumulative drug dose	"Drug_exposure"	262,770	5630
Administered medication	"Drug_exposure"	262,770	5630
Prescribed medication	"Drug_exposure"	341,986	5639
Movements between units	"Visit_detail"	47,817	6791
Clinical observation	"Measurement"	6,736,745	5973
Laboratory observation	"Measurement"	3,392,873	6001
Limitation of life-sustaining treatment	"Observation"	1298	1142
Procedure	"Procedure_occurrence"	19,861	4994

[a]EHR: electronic health record.

[b]OMOP CDM: Observational Medical Outcomes Partnership Common Data Model.

[c]EN/ISO: European Norm/International Organization for Standardization.

OMOP CDM–Compliant Clinical Tables

We still do not have the final number of records in our OMOP tables, since the initial approach of adapting the preexisting API had to be replaced by the creation of the application that inserts data from the extracts into the ontologies. However, an OMOP database for a random small subset of patients was successfully created to test the queries and validate the methodology. This was performed using a locally developed Protégé plugin ("OntoLoad") that imports a set of data from a relational database into the ontologies [17]. Table 2 describes the OMOP tables that were created.

Table 2. OMOP CDM[a]-compliant clinical tables created for a random small subset of patients.

OMOP CDM table	Patients, n	Records, n
"Condition occurrence"	121	864
"Death"	110	110
"Device_exposure"	3	56
"Drug_exposure"	106	5609
"Measurement"	3	2091
"Observation"	3	195
"Observation_period"	897	897
"Person"	922	922
"Visit_detail"	250	772
"Visit_occurrence"	897	971

[a]OMOP CDM: Observational Medical Outcomes Partnership Common Data Model.

Discussion

Principal Results

This study proposes a methodology for standardizing clinical data, thus allowing its reuse without any change in the meaning of the modeled concepts. Although the focus of this paper is health research, our methodology suggests that the data be initially standardized according to EN/ISO 13606 to obtain EHR extracts with a high level of granularity that can be used for any purpose, as previous studies have suggested [20]. Afterward, its transformation to OMOP CDM–compliant tables allows its consolidation in joint repositories for research purposes.

Although EN/ISO 13606 was chosen because of the operational mechanisms it offers for data exchange, due to the flexibility and standard-agnostic nature of our methodology, there is complete independence regarding any specific standard. Thus,

by modeling ontologies of other standards and mapping them to local variables, we may, for example, carry out transformations between EN/ISO 13606, OpenEHR [21], Fast Healthcare Interoperability Resources (FHIR) [22], OMOP CDM, and Informatics for Integrating Biology and the Bedside (i2b2) [23] with the minimum use of resources and without the need for changes in the database structure. Health information standards such as EN/ISO 13606 and OpenEHR allow the modeling and formalization of clinical knowledge through their RMs and archetypes [24], and ontologies are precisely a tool for carrying out such tasks. This is what makes them ideal in the context of an implementation of a dual-model strategy, allowing the representation of concepts in the health domain, its standardization, and the storage of instantiated patient data.

Furthermore, ontologies provide several advantages for the conceptualization of entities in a domain. It explicitly represents domain knowledge, allows the application of inference processes, enables the reuse of domain knowledge, allows data aggregation, and detects new associations between concepts [17].

It is clear for us that loading normalized data onto clinical repositories (instead of ad hoc data loading) provides many benefits. It is possible to reuse the same interoperability standards used in health care, adapting them to this new paradigm [25]. This approach allows the availability of clinical data for further single- or multicenter research.

We would like to highlight the vital importance of continuous collaborative research. This study is framed within a continued line of research since 2009 between HCB, ISCIII, and H12O. In this line of collaborative research, a standardized and transparent process has been designed and implemented for obtaining standardized data models for research from EHR raw data. Hence, in the first stage, the basis for a semantically interoperable clinical information management system based on EN/ISO 13606 was defined, proving that clinical information residing in heterogeneous systems could be normalized, combined, and communicated without loss of meaning. In the second stage, a common information model that reflects the clinical process and the relationships between the clinical records components was developed. In the third stage, a normalized information model based on EN/ISO 13606 archetypes was implemented and applied to local information systems for specific clinical use cases. With this model, it is possible to construct and order information recovered from these complex systems for the exchange of integral health and social information of patients and to use it for secondary purposes.

Comparison With Prior Work

Many of the requisites of clinical data repositories for primary use are common to those for secondary use, such as normalized clinical information models, controlled terminologies, identification of actors, and contextual information. Developments carried out for primary use repositories are also profitable for secondary uses, and the progresses derived from secondary uses accelerate the advances in shared clinical records. A lot of work has been reported in this field throughout the globe in the last years, which has led to developing policies, repository models and its application in the form of competitive projects [2,26,27].

It is very usual for researchers to resort to the generation of their own data for research and its manual introduction into data management systems. It is also quite common for them to use general purpose tools, particularly spreadsheets, as data management systems [28], while there is perception of a high need of additional support for the analysis of high volumes of data. This represents a significant problem, since these applications cannot guarantee the consistency of data, and they present difficulties for sharing and consolidating data and a limited capability of data exploitation.

Different methodologies have been proposed to create OMOP repositories from raw data. Some approaches are based on a simple mapping of local variables to their OMOP CDM counterparts, an alignment of vocabularies using the Athena tool provided by OHDSI and an ETL process through SQL scripts [29]. Other authors have proposed transforming source data to RDF, carrying out a semantic mapping (in some cases, using an ontological representation of OMOP CDM), and loading it to a data store [30,31].

Likewise, other standard-agnostic approaches have been reported in the literature. The ongoing INFOBANCO project of the Madrid Region [32] seeks to create a platform for the management, persistence, exchange, and reuse of health data focused on applying each health information standard for the purpose it was intended to, offering multiple interoperability and exploitation services suited for specific use cases [24]. Furthermore, the 3-pillar strategy of the Swiss Personalized Health Network [33] pursues a semantically interoperable clinical data landscape based on a multidimensional encoding of concepts, an RDF-based storage and transport of the instances of these concepts and a conversion of RDF to any target data model.

Strengths and Limitations

This study has many strengths that are worth mentioning. On the one hand, it describes a real-world collaborative effort between 3 health care institutions in Spain to model, share, and consolidate standardized patient data. Furthermore, the standard-agnostic nature of the proposed methodology leads to a significant scalability, allowing transformation between different health information standards and common data models. The software used in our methodology (LinkEHR, Protégé, and Liferay) either have a free version or are open source, which make them accessible to low-income areas and institutions with limited funding for interoperability projects.

We must also mention the limitations of this study. First of all, the ontology-based clinical repository used in our institution was developed throughout many years, and it might not be a suitable approach for institutions that seek a rapid implementation of a methodology. This can limit the external validity of the study. Moreover, since the tool to insert data from standardized extracts into the ontologies is not ready yet, we still have not completed the creation of OMOP CDM tables. However, an OMOP CDM database for a small subset of

patients was successfully created to test the queries and validate the methodology.

Next Steps

The MIU team at HCB is working on creating the ontological representation of different health information standards (FHIR and OpenEHR) and CDMs (i2b2, International Cancer Genome Consortium Accelerating Research in Genomic Oncology (ICGC Argo) [34], and Clinical Data Interchange Standards Consortium (CDISC) [35]). This will extend the current metamodel and allow us to carry out multistandard transformations, which will also help us compare the performance of such standards for different scenarios.

Conclusions

Semantic interoperability plays a very important role within HISs, providing meaning and clinical context to the clinical information and allowing for better clinical decision-making and research. This study has demonstrated that ontologies constitute a valuable approach for knowledge representation and standardization of health information in a standard-agnostic manner. With the proposed methodology, institutions can go from local raw data to standardized, semantically interoperable EN/ISO 13606 and OMOP repositories.

Acknowledgments

This study is also framed within the Spanish Secretary of State for Telecommunications and Digital Infrastructure's "Plan de Impulso de las Tecnologías del Lenguaje" (Plan TL). We would like to thank the Instituto de Salud Carlos III (ISCIII), VeraTech For Health, and Barcelona Supercomputing Center for their collaboration on this project. This work was supported by the ISCIII and cofunded by the European Union (grant PI18/00890, PI18/00981, and PI18CIII/00019).

Conflicts of Interest

None declared.

References

1. Jungkunz M, Köngeter A, Mehlis K, Winkler EC, Schickhardt C. Secondary use of clinical data in data-gathering, non-interventional research or learning activities: definition, types, and a framework for risk assessment. J Med Internet Res 2021;23(6):e26631 [FREE Full text] [doi: 10.2196/26631] [Medline: 34100760]
2. Safran C, Bloomrosen M, Hammond WE, Labkoff S, Markel-Fox S, Tang PC, Expert Panel. Toward a national framework for the secondary use of health data: an American Medical Informatics Association white paper. J Am Med Inform Assoc 2007;14(1):1-9 [FREE Full text] [doi: 10.1197/jamia.M2273] [Medline: 17077452]
3. Robertson ARR, Nurmatov U, Sood HS, Cresswell K, Smith P, Sheikh A. A systematic scoping review of the domains and innovations in secondary uses of digitised health-related data. J Innov Health Inform 2016;23(3):611-619 [FREE Full text] [doi: 10.14236/jhi.v23i3.841] [Medline: 28059695]
4. Ohmann C, Banzi R, Canham S, Battaglia S, Matei M, Ariyo C, et al. Sharing and reuse of individual participant data from clinical trials: principles and recommendations. BMJ Open 2017;7(12):e018647 [FREE Full text] [doi: 10.1136/bmjopen-2017-018647] [Medline: 29247106]
5. Hey T. The fourth paradigm—data-intensive scientific discovery. In: Kurbanoğlu S, Al U, Erdoğan PL, Tonta Y, Uçak N, editors. E-Science and Information Management. IMCW 2012. Communications in Computer and Information Science, vol 317. Berlin: Springer; 2012.
6. Pedrera-Jiménez M, García-Barrio N, Rubio-Mayo P, Tato-Gómez A, Cruz-Bermúdez JL, Bernal-Sobrino JL, et al. TransformEHRs: a flexible methodology for building transparent ETL processes for EHR reuse. Methods Inf Med 2022;61(S 02):e89-e102 [FREE Full text] [doi: 10.1055/s-0042-1757763] [Medline: 36220109]
7. Health informatics—electronic health record communication—Part 1: Reference model. ISO. URL: https://www.iso.org/standard/67868.html [accessed 2023-02-08]
8. Health informatics—electronic health record communication—Part 2: Archetype interchange specification. ISO. URL: https://www.iso.org/standard/62305.html [accessed 2023-02-08]
9. ISO 13940:2015—Health informatics—system of concepts to support continuity of care. ISO. URL: https://www.iso.org/standard/58102.html [accessed 2022-06-04]
10. OHDSI. The Book of OHDSI: Observational Health Data Sciences and Informatics. North Bethesda, MD: OHDSI; 2019.
11. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. Stud Health Technol Inform 2015;216:574-578 [FREE Full text] [Medline: 26262116]
12. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc 2012;19(1):54-60 [FREE Full text] [doi: 10.1136/amiajnl-2011-000376] [Medline: 22037893]

13. Garza M, Del Fiol G, Tenenbaum J, Walden A, Zozus MN. Evaluating common data models for use with a longitudinal community registry. J Biomed Inform 2016;64:333-341 [FREE Full text] [doi: 10.1016/j.jbi.2016.10.016] [Medline: 27989817]
14. Maldonado JA, Moner D, Boscá D, Fernández-Breis JT, Angulo C, Robles M. LinkEHR-Ed: a multi-reference model archetype editor based on formal semantics. Int J Med Inform 2009;78(8):559-570. [doi: 10.1016/j.ijmedinf.2009.03.006] [Medline: 19386540]
15. Maggio LA, Stranack K. Understanding Creative Commons. Acad Med 2020;95(2):322. [doi: 10.1097/ACM.0000000000003031] [Medline: 31599759]
16. Lozano-Rubí R, Muñoz Carrero A, Serrano Balazote P, Pastor X. OntoCR: a CEN/ISO-13606 clinical repository based on ontologies. J Biomed Inform 2016;60:224-233 [FREE Full text] [doi: 10.1016/j.jbi.2016.02.007] [Medline: 26911524]
17. Lozano-Rubí R. A Metamodel for Clinical Data Integration: Basis for a New EHR Model Driven by Ontologies. 2017. URL: https://www.tdx.cat/bitstream/handle/10803/399855/rlr1de1.pdf?sequence=1 [accessed 2023-02-08]
18. OWL Web Ontology Language Reference. OWL. URL: http://www.w3.org/TR/owl-features [accessed 2022-05-13]
19. Musen MA, Protégé Team. The Protégé project: a look back and a look forward. AI Matters 2015;1(4):4-12 [FREE Full text] [doi: 10.1145/2757001.2757003] [Medline: 27239556]
20. Pedrera-Jiménez M, García-Barrio N, Cruz-Rojo J, Terriza-Torres AI, López-Jiménez EA, Calvo-Boyero F, et al. Obtaining EHR-derived datasets for COVID-19 research within a short time: a flexible methodology based on detailed clinical models. J Biomed Inform 2021;115:103697 [FREE Full text] [doi: 10.1016/j.jbi.2021.103697] [Medline: 33548541]
21. Kalra D, Beale T, Heard S. The openEHR Foundation. Stud Health Technol Inform 2005;115:153-173. [Medline: 16160223]
22. Ayaz M, Pasha MF, Alzahrani MY, Budiarto R, Stiawan D. The fast health interoperability resources (FHIR) standard: systematic literature review of implementations, applications, challenges and opportunities. JMIR Med Inform 2021;9(7):e21929 [FREE Full text] [doi: 10.2196/21929] [Medline: 34328424]
23. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc 2010;17(2):124-130 [FREE Full text] [doi: 10.1136/jamia.2009.000893] [Medline: 20190053]
24. Pedrera-Jiménez M, Spanish Expert Group on EHR standards, Kalra D, Beale T, Muñoz-Carrero A, Serrano-Balazote P. Can OpenEHR, ISO 13606 and HL7 FHIR work together? An agnostic perspective for the selection and application of EHR standards from Spain. TechRxiv. Preprint posted online on May 25, 2022 [FREE Full text] [doi: 10.36227/techrxiv.19746484]
25. Haarbrandt B, Tute E, Marschollek M. Automated population of an i2b2 clinical data warehouse from an openEHR-based data repository. J Biomed Inform 2016;63:277-294 [FREE Full text] [doi: 10.1016/j.jbi.2016.08.007] [Medline: 27507090]
26. Pathak J, Bailey KR, Beebe CE, Bethard S, Carrell DS, Chen PJ, et al. Normalization and standardization of electronic health records for high-throughput phenotyping: the SHARPn consortium. J Am Med Inform Assoc 2013;20(e2):e341-e348 [FREE Full text] [doi: 10.1136/amiajnl-2013-001939] [Medline: 24190931]
27. Cimino JJ, Ayres EJ. The clinical research data repository of the US National Institutes of Health. Stud Health Technol Inform 2010;160(Pt 2):1299-1303 [FREE Full text] [Medline: 20841894]
28. Anderson NR, Lee ES, Brockenbrough JS, Minie ME, Fuller S, Brinkley J, et al. Issues in biomedical research data management and analysis: needs and barriers. J Am Med Inform Assoc 2007;14(4):478-488 [FREE Full text] [doi: 10.1197/jamia.M2114] [Medline: 17460139]
29. Paris N, Lamer A, Parrot A. Transformation and evaluation of the MIMIC database in the OMOP common data model: development and usability study. JMIR Med Inform 2021;9(12):e30970 [FREE Full text] [doi: 10.2196/30970] [Medline: 34904958]
30. Pacaci A, Gonul S, Sinaci AA, Yuksel M, Laleci Erturkmen GB. A semantic transformation methodology for the secondary use of observational healthcare data in postmarketing safety studies. Front Pharmacol 2018;9:435 [FREE Full text] [doi: 10.3389/fphar.2018.00435] [Medline: 29760661]
31. Sun H, Depraetere K, De Roo J, Mels G, De Vloed B, Twagirumukiza M, et al. Semantic processing of EHR data for clinical research. J Biomed Inform 2015;58:247-259 [FREE Full text] [doi: 10.1016/j.jbi.2015.10.009] [Medline: 26515501]
32. infobank. Comunidad de Madrid. URL: https://cpisanidadcm.org/infobanco/ [accessed 2023-02-08]
33. Gaudet-Blavignac C, Raisaro JL, Touré V, Österle S, Crameri K, Lovis C. A national, semantic-driven, three-pillar strategy to enable health data secondary usage interoperability for research within the Swiss personalized health network: methodological study. JMIR Med Inform 2021;9(6):e27591 [FREE Full text] [doi: 10.2196/27591] [Medline: 34185008]
34. Accelerating research in genomic oncology. ICGC ARGO. URL: https://www.icgc-argo.org/ [accessed 2023-02-08]
35. Standards. CDISC. URL: https://www.cdisc.org/standards [accessed 2023-02-08]

Abbreviations

AM: archetype model
API: application programming interface
BSC: Barcelona Supercomputing Center

CDISC: Clinical Data Interchange Standards Consortium
EHR: electronic health record
EN/ISO 13606: European Norm/International Organization for Standardization
ER: emergency room
ETL: extract, transform, and load
FHIR: Fast Healthcare Interoperability Resources
H12O: Hospital 12 de Octubre
HCB: Hospital Clínic de Barcelona
HIS: health information system
i2b2: Informatics for Integrating Biology and the Bedside
ICD-10-CM: International Classification of Diseases 10—Clinical Modification
ICGC Argo: International Cancer Genome Consortium Accelerating Research in Genomic Oncology
ISCIII: Instituto de Salud Carlos III
LOINC: Logical Observation Identifiers Names and Codes
MIU: Medical Informatics Unit
OMOP CDM: Observational Medical Outcomes Partnership Common Data Model
OWL: Web Ontology Language
OWL-DB: Web Ontology Language database
RDF: Resource Description Framework
RM: reference model
SNOMED CT: Systematized Nomenclature of Medicine—Clinical Terms

Edited by A Benis; submitted 23.11.22; peer-reviewed by R Sánchez de Madariaga, P Azevedo Marques; comments to author 20.12.22; revised version received 28.12.22; accepted 05.01.23; published 08.03.23.

Please cite as:
Frid S, Pastor Duran X, Bracons Cucó G, Pedrera-Jiménez M, Serrano-Balazote P, Muñoz Carrero A, Lozano-Rubí R
An Ontology-Based Approach for Consolidating Patient Data Standardized With European Norm/International Organization for Standardization 13606 (EN/ISO 13606) Into Joint Observational Medical Outcomes Partnership (OMOP) Repositories: Description of a Methodology
JMIR Med Inform 2023;11:e44547
URL: https://medinform.jmir.org/2023/1/e44547
doi:10.2196/44547
PMID:36884279

Original Paper

Data-Driven Identification of Unusual Prescribing Behavior: Analysis and Use of an Interactive Data Tool Using 6 Months of Primary Care Data From 6500 Practices in England

Lisa EM Hopcroft[1], PhD; Jon Massey[1], PhD; Helen J Curtis[1], DPhil; Brian Mackenna[1], MPharm; Richard Croker[1], MSc; Andrew D Brown[1], MPharm, MPsych; Thomas O'Dwyer[1], BSc; Orla Macdonald[2], MPharm; David Evans[1], MPhil; Peter Inglesby[1], MPhil; Sebastian CJ Bacon[1], BA; Ben Goldacre[1], MRCPsych; Alex J Walker[1], PhD

[1]Bennett Institute for Applied Data Science, Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, United Kingdom
[2]Oxford Health NHS Foundation Trust, Oxford, United Kingdom

Corresponding Author:
Alex J Walker, PhD
Bennett Institute for Applied Data Science
Nuffield Department of Primary Care Health Sciences
University of Oxford
Radcliffe Primary Care Building
32 Woodstock Road
Oxford, OX2 6GG
United Kingdom
Phone: 44 01865617855
Fax: 44 01865617855
Email: alex.walker@phc.ox.ac.uk

Abstract

Background: Approaches to addressing unwarranted variation in health care service delivery have traditionally relied on the prospective identification of activities and outcomes, based on a hypothesis, with subsequent reporting against defined measures. Practice-level prescribing data in England are made publicly available by the National Health Service (NHS) Business Services Authority for all general practices. There is an opportunity to adopt a more data-driven approach to capture variability and identify outliers by applying hypothesis-free, data-driven algorithms to national data sets.

Objective: This study aimed to develop and apply a hypothesis-free algorithm to identify unusual prescribing behavior in primary care data at multiple administrative levels in the NHS in England and to visualize these results using organization-specific interactive dashboards, thereby demonstrating proof of concept for prioritization approaches.

Methods: Here we report a new data-driven approach to quantify how "unusual" the prescribing rates of a particular chemical within an organization are as compared to peer organizations, over a period of 6 months (June-December 2021). This is followed by a ranking to identify which chemicals are the most notable outliers in each organization. These outlying chemicals are calculated for all practices, primary care networks, clinical commissioning groups, and sustainability and transformation partnerships in England. Our results are presented via organization-specific interactive dashboards, the iterative development of which has been informed by user feedback.

Results: We developed interactive dashboards for every practice (n=6476) in England, highlighting the unusual prescribing of 2369 chemicals (dashboards are also provided for 42 sustainability and transformation partnerships, 106 clinical commissioning groups, and 1257 primary care networks). User feedback and internal review of case studies demonstrate that our methodology identifies prescribing behavior that sometimes warrants further investigation or is a known issue.

Conclusions: Data-driven approaches have the potential to overcome existing biases with regard to the planning and execution of audits, interventions, and policy making within NHS organizations, potentially revealing new targets for improved health care service delivery. We present our dashboards as a proof of concept for generating candidate lists to aid expert users in their interpretation of prescribing data and prioritize further investigations and qualitative research in terms of potential targets for improved performance.

(JMIR Med Inform 2023;11:e44237) doi:10.2196/44237

KEYWORDS

dashboard; data science; EHR; electronic health records; general practice; outliers; prescribing; primary care

Introduction

There is recognition that evidence-based decision-making in the National Health Service (NHS) in England is critical to maintaining standards of care while reducing NHS spending [1] and the UK government has recently consulted on wide-ranging plans to "digitize, connect, and transform the health and care sector," with a key priority being data-driven innovation. Flagship initiatives such as Getting It Right First Time [2] and RightCare [3] focus on identifying and addressing unwarranted variation in the NHS. Such initiatives can be limited in their scope in that the "data-driven" element of the work often focuses on assessing performance relative to recommendations that are defined prospectively rather than employing hypothesis-free data-driven methodologies to make objective assessments as to where opportunities for improvement might exist.

Monthly prescription data for every general practice in England has been made available to the public since 2010, via the NHS Business Services Authority [4]. This data set includes product and month of prescribing, the number of items prescribed and the total quantity, making it very amenable to detailed analysis for the purposes of original research [5-9] and systematic audits and reviews [10,11]. These data, made navigable via interactive dashboards [12,13], are commonly used by NHS staff—in particular, medicines optimization (MO) teams—to monitor performance on key prescribing indicators, compare performance to peer organizations, inform prioritization of work streams, estimate the impact and feasibility of interventions, or create customized outputs according to local priorities. Mining these data systematically for unusual prescribing behavior could help identify where service delivery improvements are possible in the absence of human bias or expectation. Such "unbiased" or "hypothesis-free" approaches might aid local decision makers when designing appropriate interventions and policies.

The value of deploying systematic analyses to large prescribing data sets has been demonstrated elsewhere. Using regional prescribing claims data from Germany, researchers were able to identify practices prescribing more "third-level" medications (ie, not first- or second-line treatments) than expected using funnel plots and mixed effects models [14]. Our own group has successfully deployed similar outlier detection methodology on a national scale to show that the prescribing of 2 antipsychotic drugs, in very limited use nationally, is concentrated in 2 small geographic regions of England [15]. More complex outlier analysis of wholesale codeine time series data has identified significant shifts in supply occurring around the time of regulatory changes (specifically, the up-scheduling of low-dose codeine products from over-the-counter to prescription-only) [16].

We run OpenPrescribing [13], a website that allows public interrogation and visualization of primary care prescription data at multiple administrative levels in the NHS in England. We have previously deployed novel methodologies to identify changes over time in any one of the 80 measures implemented in OpenPrescribing, providing monthly alerts to notify practitioners when their prescribing rates deviate from the norm and may require clinician attention [17]. These measures have been selected on the basis of clear guidance being available from health authorities and are subject to initial and continuing review by clinicians, pharmacists, and epidemiologists. OpenPrescribing has 20,000 unique users every month and thousands of subscribers to our innovative organization email alerts service [17].

We set out to develop new hypothesis-blind data science techniques to identify unusual prescribing behavior, thereby providing proof of concept for such an analysis and illustrating potential opportunities for service improvement. Using this approach, we have no hypothesis with regard to where interesting patterns might be found (ie, which clinical area or which organization), we only have an expectation of what would constitute an interesting pattern in the data. We applied this methodology to 6 months of national prescribing data to identify outliers at multiple administrative levels of the NHS in England during that time period, presenting the most extreme outliers in each organization for the consideration of expert users to prioritize for further review, qualitative research, and interpretation within the local context.

Methods

Study Design

Prescribing practice was analyzed by conducting a retrospective cohort study using prescribing data from all English NHS general practices, primary care networks (PCNs), clinical commissioning groups (CCGs), and sustainability and transformation partnership (STPs; Textbox 1).

Textbox 1. National Health Service (NHS) England administrative organizations.

- Primary care in England is delivered by individual general practices, with one or more general practitioners. Almost all (>99%) practices are grouped together with other local primary care provision to form primary care networks, typically representing 30,000-50,000 people [18]. In the study period, practices were also grouped together into Clinical Commissioning Groups (CCGs) which were clinically-led organizations responsible for the commission of primary (and secondary) care in a geographical region [19]. As of April 2021, there were 106 CCGs in England. In the study period, CCGs were clustered into Sustainability and Transformation Partnerships (STPs) [20]. As of May 2020, there were 42 STPs in England. In July 2022, CCGs and STPs were replaced with 42 integrated care boards (ICBs), though the data used in this study predates this change.
- Also important in this architecture are the medicines optimization teams—NHS staff who provide expert advice with regard to medicines commissioning, finance, and safety [21]. Medicines optimization teams have historically operated at the level of CCGs (or their sub-ICB replacements) but are increasingly operating at the broader level of ICBs.

Data Source

Data for the period June 1, 2021, to December 1, 2021, were extracted from the OpenPrescribing database; this 6-month study period was used so as to smooth out short-term fluctuations (by aggregating multiple months of data) while keeping to a relatively recent time frame (so that the data remain relevant). OpenPrescribing imports openly accessible prescribing data from the large, monthly files published by the NHS Business Services Authority, which contain data on cost and items prescribed for each month for every typical general practice and CCG in England, dating back to mid-2010 [4,22]. These data are published only at the level of organization; patient-level data are not made available. Detailed methods for the creation of OpenPrescribing, including data management, aggregation, and cleaning, are available elsewhere [23]. The monthly prescribing data sets contain 1 row for each different medication and dose in each prescribing organization in NHS primary care in England, describing the number of items (ie, prescriptions issued) and the total cost. These data are sourced from community pharmacy claims data and, therefore, contain all items that were dispensed. All available prescribing data were extracted for institutions identified as "typical" general practices; all other organizations, such as prisons or specialist community clinics, were excluded using NHS Digital organization data [24]. We limited our analysis to the 2369 chemicals from chapters 1-15 of the British National Formulary (BNF) to exclude chapters not following a chemical and subparagraph structure, those which largely cover nonmedicinal products such as dressings (see Textbox 2 for further information regarding prescribing terminology).

Textbox 2. Prescribing terminology.

- The public prescribing data made available by the National Health Service Business Services Authority uses a pseudo-British National Formulary (BNF) classification. The most granular level of data is at "presentation" level, which includes information on the prescription medicine, brand, strength, and formulation. This data can then be grouped using the pseudo-BNF hierarchy, using products, chemical substances, subparagraphs, paragraphs, sections, and chapters, with decreasing specificity. Chapters are defined according to body system, for example, gastrointestinal system, cardiovascular system, and respiratory system.
- "Chemical" in this context refers to the standard International Nonproprietary Name (INN) for the active constituent of the medicine and does not include any further specification by preparation, dose, or brand. BNF subparagraphs can be used to identify groups of chemicals belonging to the same class.
- The majority of chemicals have all available preparations included in a single chemical definition; for example, all atorvastatin preparations (including liquid and tablets) are included in 2.12: Cardiovascular system—lipid-regulating drugs (chemical code: 0212000B0).
- However, there are some instances where the same chemical is used in different body systems with system-specific presentations, and therefore the same INN will appear multiple times in a chapter that most reflects its use. For example, the INN dexamethasone appears in 3 separate chapters within the pseudo-BNF hierarchy and therefore will have separate chemical groupings:

6.3: Endocrine system—corticosteroids (endocrine), which include oral and parenteral preparations (chemical code: 0603020G0)

11.4: Eye—corticosteroids and other anti-inflammatory preparations, which include ocular preparations (chemical code: 1104010I0)

12.1 Ear, Nose, and Oropharynx—drugs acting on the ear, which ear preparations (chemical code: 12101050)

Outlier Detection

We were interested in detecting outliers with regard to *chemicals* (see Textbox 2 for further information). We first calculate a prescription rate for each chemical in each practice; specifically, we calculate the number of prescriptions containing our chemical of interest and divide this by the number of prescriptions containing chemicals of the same BNF subparagraph, for example, all statin prescriptions as a proportion of all lipid-regulating drugs. This captures the prescribing rate for the chemical of interest as compared to all drugs in the same class in a single practice. This ratio is calculated across all practices, and the mean and SD are calculated. The ratios in each practice are then reexpressed as *z* scores using this mean and SD. A *z* score is the number of SDs that a given data point is away from the mean. The *z* scores are used to rank all chemicals within a practice in terms of their outlier status (the most extreme outliers occupying the top and bottom of this ranked list).

This process is repeated at 3 higher administrative levels—STP, CCG, and PCN—to generate the equivalent ranked list of

prescribing outliers for these larger organizations. Results for all 4 administrative levels are presented, as each organization retains some decision-making power with regard to prescribing. At the practice or PCN level this will be individual or group general practitioner decisions based on their practice population, but MO teams at the STP and CCG level will also monitor prescribing behavior to inform prescribing policy (and formulary) for these wider geographic regions.

Visualization of Organization-Level Results

An interactive dashboard has been created at OpenPrescribing [25] for each organization, where data describing 20 of the most extreme outliers are summarized as follows: 10 where prescribing in the organization is *higher* than other peer organizations, and 10 where prescribing in the organization is *lower* than other peer organizations. Tables are provided for both sets, which summarize the following values for each chemical: *Chemical Items* and *Subparagraph Items* are the number of prescriptions for the chemical and BNF subparagraph, respectively; *Ratio* is the *Chemical Items* as a proportion of *Subparagraph Items* for the chemical in the organization of interest; *Mean* and *SD* summarize this ratio over all organizations; and *z score* is the *Ratio* reexpressed as a *z* score. This same information is described visually by a density plot (provided in the Multimedia Appendices), where the distribution of ratios across all organizations is captured by a blue line, with the ratio for the organization of interest indicated by a vertical red line. Densities are generated using the Seaborn kdeplot() function, setting the bandwidth for smoothing as suggested by Scott [26].

User Feedback

Links to early prototypes were shared directly with a group of interested clinicians and pharmacists by email. Any feedback gained was used to inform the iterative development of the tool and proposed visualizations of the results. Further to this, the tool was shared more widely (via Twitter), and formal feedback was collected via a Google form (Textbox 3). Additional unstructured feedback was compiled from direct emails and mentions on social media.

Textbox 3. Outlier detection feedback form.

Respondent details:

- Email. Free text
- Which organization's report are you giving feedback on? Free text
- Please describe your relationship to the organization (eg, doctor, practice nurse, or commissioner). Free text

Understandability:

- Does this report make sense to you? Yes or No
- Any further comments on the understandability of the report(s). Free text

Interest:

- Is it interesting? Yes or No
- Any further comments on the interestingness of the report(s). Free text

Utility:

- Is it useful? Yes or No
- Any further comments on the usefulness of the report(s). Free text

Individual items:

- Thinking about where your prescribing is higher than most, please describe any observations you have on any individual items. Free text
- Thinking about where your prescribing is lower than most, please describe any observations you have on any individual items. Free text

Improvements:

- What, if anything, would you change about the report(s)? Free text

NHS Devon CCG Case Study Details

RC (who, in addition to his role at the Bennett Institute, is also Deputy Director for MO at NHS Devon) emailed a link to the dashboard containing sparkline graphs for NHS Devon to pharmacist colleagues in his MO team. These graphs provided new insights to the team, which would have been impractical to achieve using existing data analysis workflows (eg, custom queries in OpenPrescribing or ePACT2). The MO team met to discuss what the causes behind the deviation might be in each case. Where it could not be determined that there was a clinically justifiable reason for being an outlier, the MO team gathered further relevant prescribing data from routine sources such as OpenPrescribing, ePACT2, and PrescQIPP. This allowed deeper exploration of prescribing patterns related to the outlier chemical (eg, trends over time and the rate at which alternative medications were prescribed). The MO team continues to

investigate these data to decide whether an intervention is appropriate.

Software and Reproducibility

Data management was performed using Python 3.8.1 and Google BigQuery, with analysis carried out using Python. Code for data management and analysis is archived on the internet [27] and dashboards are available on the OpenPrescribing website [25].

Patient and Public Involvement

We publicized this tool via social media and actively sought feedback from interested health care professionals and members of the public to inform its iterative development via a survey (see User Feedback section above). We will continue to seek and consider feedback via these same channels as the tool is developed. We have developed a publicly available website [13] through which we invite any patient or member of the public to contact us regarding this study or the broader OpenPrescribing project.

Results

Outlier Detection

We developed interactive dashboards for every practice in England to highlight unusual prescribing. The outlying chemicals (ie, the 10 chemicals ranked highest and 10 chemicals ranked lowest by *z* score) identified using our methodology are described in Table 1. Both counts of unique chemicals and summary statistics of *z* scores are provided at each of the 4 administrative levels. Those outlying chemicals that are "higher than most" will all have positive *z* and as such are summarized using the maximum, median, Q1 and Q3; similarly, outlying chemicals that are "lower than most" will have negative *z* scores, and are summarized using the minimum, median, Q1 and Q3. A measure of the variation in the *z* score amongst all organizations at the same administrative level can be obtained by calculating the Inter Quartile Range (IQR), defined as Q3-Q1.

Table 1. Summary statistics for z scores calculated for outlying chemicals across the 4 administrative levels. Outlying chemicals are those occurring in the top 10 (ie, "Higher than most") or bottom 10 (ie, "Lower than most") by z score in at least one organization at the corresponding administrative level.

Organization type	Unique chemicals, n	Higher than most			Lower than most		
		Maximum	Median	Q1-Q3	Minimum	Median	Q1-Q3
STP[a] (n=42)	680	6.33	5.42	4.6-6.24	−6.33	−2.35	−2.87–−2.08
CCG[b] (n=106)	1138	10.20	5.79	4.59-7.77	−10.20	−2.30	−2.81–−1.99
PCN[c] (n=1257)	1416	2528.09	5.28	4.17-7.56	−159.77	−2.18	−2.67–−1.9
Practice (n=6476)	1346	6825.50	5.23	3.93-7.77	−307.23	−2.08	−2.57–−1.76

[a]STP: sustainability and transformation partnership.

[b]CCG: clinical commissioning group.

[c]PCN: primary care network.

While the median values for the "higher than most" outlying chemicals are similar, the IQR (Q3-Q1) values demonstrate that variation between peer organizations decreases with the size of the organization; the least amount of variation is observed between STPs, and the most amount of variation is observed between practices. More outlying chemicals are identified in smaller organizations (PCNs and practices). With regard to outlying chemicals identified as being prescribed at lower rates compared to peer organizations, both the median and IQR of the *z* scores are very similar across all organization types. For both sets of outlying chemicals, the most extreme outliers occur further away from the mean as the organization size decreases; the maximum value for the "higher than most" outlying chemicals *increases* with the size of the organization and the minimum value for the "lower than most" outlying chemicals *decreases* with the size of the organization. The *z* scores for "higher than most" outlying chemicals are more extreme than the "lower than most" outlying chemicals in all organization types.

Organization-Level Results Visualization: Case Study of NHS Devon CCG

NHS Devon CCG is the fifth largest CCG in England, commissioning health care for 1.2 million people in the southwest of England. The top 10 chemicals that are prescribed at *higher* rates here compared to other CCGs are shown in the top portion of Table 2, while the top 10 chemicals that are prescribed at *lower* rates are shown in the bottom portion of Table 2 (a listing of the specific products and a sparkline plot, showing graphically where the ratio value for this CCG occurs in the context of the same ratio in all CCGs, are provided in Multimedia Appendix 1). These prescribing outliers for this CCG have been reviewed by the local MO team, to provide likely explanations for the outlier prescribing.

Table 2. The outlier detection dashboard for NHS[a] Devon clinical commissioning group (CCG)[b].

BNF[c] chemical (number of products)	Chemical items, n	BNF subparagraph	Subparagraph items, n	Ratio	Mean	SD	*z* score
Prescribing where NHS Devon CCG is *higher* than most							
Levobupivacaine hydrochloride (1)	130	Local anesthetics	20,482	0.01	0	0	10.2
Gripe mixtures (1)	1	Sodium bicarbonate	56	0.02	0	0	8.34
Gluten free pastas (3)	4	Foods for special diets	9199	0	0	0	7.97
Epoetin zeta (1)	2	Hypoplastic, hemolytic, and renal anemias	18	0.11	0	0.01	7.52
Flumetasone pivalate (1)	333	Otitis externa	19,724	0.02	0	0	6.98
Gluten free or wheat free cereals (1)	2	Foods for special diets	9199	0	0	0	6.81
Levofloxacin (2)	1372	Quinolones	4754	0.29	0.05	0.04	6.61
Liquefied phenol (1)	1	Phenolics	3	0.33	0.01	0.06	5.83
Ruxolitinib (1)	2	Other antineoplastic drugs	1494	0	0	0	5.10
Ferrous gluconate (1)	10,437	Oral iron	90,095	0.12	0.03	0.02	3.66
Prescribing where NHS Devon CCG is *lower* than most							
Sodium bicarbonate (3)	55	Sodium bicarbonate	56	0.98	1	0	–8.34
Ciprofloxacin (6)	2989	Quinolones	4754	0.63	0.86	0.05	–4.22
Dexamethasone (2)	13,061	Otitis externa	19,724	0.66	0.79	0.05	–2.71
Fexofenadine hydrochloride (6)	33,711	Antihistamines	169,747	0.20	0.36	0.07	–2.33
Oral rehydration salts (8)	1942	Oral sodium and water	5010	0.39	0.65	0.11	–2.27
Betamethasone esters (12)	1672	Topical corticosteroids	141,063	0.01	0.03	0.01	–2.20
Fusidic acid (1)	359	Antibacterials	13,283	0.03	0.08	0.03	–2.12
Senna (9)	39,769	Stimulant laxatives	110,838	0.36	0.55	0.1	–2.03
Ticagrelor (3)	2285	Antiplatelet drugs	467,104	0	0.02	0.01	–2.02
Lactulose (2)	19,621	Osmotic laxatives	127,773	0.15	0.28	0.06	–1.89

[a]NHS: National Health Service.

[b]"The results of our outlier detection methodology are provided as interactive dashboards; here, the 10 chemicals where prescribing in National Health Service (NHS) Devon CCG is higher than most and the 10 chemicals where prescribing in NHS Devon CCG is lower than most, are presented. British National Formulary (BNF) chemical is the chemical of interest (number of products indicates how many products are represented by the BNF chemical). Chemical items provide the number of prescribing items containing this chemical. BNF subparagraph is the BNF subparagraph to which the chemical belongs, and subparagraph items is the number of prescribing items containing an item belonging to this BNF subparagraph. Ratio, Mean, SD, and *z* score place the chemical items count in the context of the subparagraph items count as described in the Methods section.

[c]BNF: British National Formulary.

Focusing on the results for flumetasone pivalate, we can see that 1.7% (n=333) of the 19,724 "Otitis externa" items contain flumetasone pivalate and that this is 6.98 SDs above the mean for all CCGs (the sparkline plot provided in Multimedia Appendix 1 demonstrates visually where this 1.7% falls [red line] in the distribution across all CCGs [blue line]).

Several of the chemicals prescribed more often in NHS Devon CCGs than other CCGs are defined as first-line treatments in local formularies, for example, flumetasone pivalate [28] and levofloxacin [29]. Corresponding patterns of underprescribing can be seen in the "lower than most" results table for similar chemicals, specifically, ciprofloxacin (an alternative to levofloxacin) and dexamethasone (an alternative to flumetasone pivalate).

The lower prescribing rates for fusidic acid reflect a change in this CCG to prescribe this chemical by specialist recommendation only [30], due to rising costs [31] and a narrow spectrum of action. The lower rates of prescribing for senna and lactulose are also likely due to a formulary shift in this CCG toward macrogols [32]. Finally, the low prescribing rate of betamethasone esters is also expected as these chemicals are nonformulary in this CCG [33].

This dashboard also demonstrates a valid use for low-number results. Gluten-free pastas and cereals—something that we have previously identified as having high variability in prescribing

rates [34]—were not recommended to be prescribed by the NHS in the study period (NHS England issued advice to CCGs in November 2018 with the recommendation to restrict gluten free prescribing to bread and flour mixes [35]), so should not appear at all. The identification of this low-number outlier via our methodology has prompted further work within NHS Devon CCG to clarify how this prescription was generated and processed.

User Feedback

Through the formal Google form and direct correspondence with interested parties, we received feedback for a prototype version of the dashboard from 6 individuals. An example of this prototype is shown in Multimedia Appendix 2, showing 5 top and bottom outlying chemicals. Several respondents indicated that the results were expected (ie, results echoed internal reporting or were aligned with local prescribing policies); while this indicates that our tool is working, 1 user did question what the added value was above existing reporting. Other users stated that the tool had revealed unexpected results worthy of follow-up.

There were multiple requests to present more than the top and bottom 5 results (eg, the top and bottom 10 or 20 results) to explore the data in more detail. Users recognized that extreme outliers could be derived from very small numbers of patients or items; some requested that results with small counts be removed, though others recognized that these may be important, particularly in practices or PCNs. There was a suggestion that users could choose to have low numbers suppressed or displayed, depending on whether their focus was systemic anomalies or rogue prescriptions. There were also requests to include other data in the results, including cost and highlighting drugs on the "*Not Suitable to Prescribe*" list.

There were other requests that were more relevant to the design of the tool than the analysis itself. The feedback demonstrated that users required more information to interpret and understand the data (ie, *z* scores, ratios, means, and SDs) and that with this additional explanation, more could be made of the graphical summary. There was also a request for an improved user experience regarding navigating to practices via the drop-down sections (which could be implemented as an organizational search).

We used the most common feedback to inform further development, and the released version of the dashboards now includes the top and bottom 10 outlying chemicals and optional filtering of low numbers. To provide a clear illustration of how the dashboards changed in response to user feedback, the corresponding update for Multimedia Appendix 2 is shown in Figures S1 and S2 in Multimedia Appendix 3.

Discussion

Summary

We have developed and implemented a new hypothesis-free methodology to detect unusual or "outlier" prescribing rates of chemicals in a single organization in relation to all "peer" organizations. We have applied this methodology to 6 months of national prescribing data to quantify how typical the prescribing is for individual chemicals at multiple administrative levels (practice, PCN, CCG, and STP) over the time period. We have displayed these results via interactive dashboards. We have sought and will continue to seek user feedback to inform development and incrementally improve usability and functionality.

Summary statistics demonstrate that the number of outlying chemicals increases as the size of the organization decreases and that more extreme outliers are identified among smaller organizations, demonstrating that there is more variability in prescribing behavior among practices than there is among larger administrative organizations. The data also demonstrate, however, that outliers *do* occur when comparing larger organizations to each other. While there is less variation between STPs, the median *z* score for "higher than most" and "lower than most" outliers among STPs is 5.42 and 2.35, respectively; these *z* scores are both more than 2 SDs from the mean. The ranking of these quantifications allows us to identify the most extreme outliers in terms of prescribing behavior at each organizational level. A case study of an individual CCG (NHS Devon) demonstrated that our methodology identified prescribing patterns that aligned with local prescribing guidance, but also detected patterns that warranted further investigation. It is not appropriate to formally assess the utility of our methodology as there are many legitimate reasons that a chemical may be an outlier in a particular organization. Some of the reasons are as follows: prescribing guidance as defined by local formulary may differ from elsewhere; local prescribing policy may place responsibility for prescribing particular drugs in secondary care rather than primary care; clinicians may be reluctant to change medication for patients who are stable on a long-established medication regime (in particular the elderly or vulnerable); or there is a justified preference for other drugs in the same class. Given the complexities of interpreting these data, we present this tool as a proof of concept and starting point for NHS organizations to perform and plan internal audits rather than a definitive reporting tool.

Strengths and Weaknesses

Our approach combines a comprehensive national prescribing data set with a well-understood system for drug classification, thereby capturing the national context at high resolution and allowing the interpretation of prescribing behavior for *all* chemicals at multiple administrative levels of the NHS in England, all of which retain some decision-making power with regard to prescribing. The methods used are well established and easy to understand, readily amenable to visual presentation as graphs, and allow prioritization of results by ranking. Our approach has utility in other contexts, and repurposing it to gain a greater understanding of other NHS data (eg, hospital prescriptions) would be straightforward.

We also note some limitations. First, the calculation of *z* scores using mean and SD assumes a normal distribution. This is more likely to be the case where numbers of items prescribed are high (aggregated to STP or CCG), but may not be the case where number of items are low (aggregated to PCN or practices, or where the items are more rarely prescribed). Second, this approach can generate very large *z* scores where SDs are tight

or item numbers generally are very low. An example of this can be seen in Figure S1 in Multimedia Appendix 3; while the ratio generated by the number of prescribed items containing Sodium aurothiomalate is very low ($1/114{,}367=8.74\times10^{6}$), the tight SDs observed across the whole population of STPs translate this small value into a large *z* score. Expert users may be seeking out such results to identify very rare prescribing items (low number results did prove important in the NHS Devon case study), but they may also wish to suppress such results to focus on more commonly prescribed chemicals. To accommodate this and in line with our user feedback, we have implemented the option to show or hide counts of 5 or less. We also recognize that the process by which we have sought user feedback thus far could be prone to bias, in that specific users were targeted due to their expertise and familiarity with such tools so as to enable rapid development.

Findings in Context

This is one of a suite of tools that we are seeking to develop at OpenPrescribing, each of which captures variability with a view to leveraging further insight from the data sets to which we have access. We make extensive use of decile plots to place individual organizations into a wider context [5,6,36] and have applied algorithms to identify when those individual organizations start to deviate from the rest of their peers [17]. We have also used deciles to summarize financial data and estimate potential savings if "price-per-unit" costs were aligned with the lowest decile [7]. These methodologies all have the potential to support NHS organizations in England to guide audits, prioritize and shape new policies, and crucially assess the impact of those interventions with regard to patient care and cost savings.

Policy Implications and Interpretation

The Department of Health and Social Care consultation explicitly recognizes the value of near real-time data release and the potential of data-driven insights to guide targeted policy making [37]. The methodology described here contributes toward that key priority by exposing specific patterns in data that warrant attention that may have otherwise been obscured. We do not advocate that our approach be used in isolation, but rather as a starting point for expert users to interpret within the local context and make evidence-based decisions about priorities and planning. By updating these dashboards on a regular basis, we hope to provide decision makers with near real-time feedback so as to monitor performance and respond quickly when necessary. Comprehensive coverage of the opportunities and challenges that exist in encouraging widespread adoption of these approaches across the NHS in England is provided in the Goldacre Review [38].

Future Research

Areas for further research include the implementation of a systematic and unbiased approach to collecting and inviting user feedback, enhancing results output as determined by ongoing user feedback (eg, new functionality, information, or visualizations), updating the dashboards in line with recent structural changes to the NHS in England (specifically, Integrated Care Boards replacing STPs), and consulting with patient and public involvement and engagement groups to maximize value for the patient community. The long-term aim is to incorporate regular updates as part of an organization's page on the OpenPrescribing website; the frequency of these updates (annual vs monthly) and the extent to which historical dashboards would be available for each organization have yet to be determined but would be a focus of the enhanced user consultation described above. Ultimately, our aim would be to provide organization specific alerts to notify staff where prescribing behavior appears to be different to their peers.

Conclusions

Capturing the variability in prescribing rates among peer organizations permits the hypothesis-free identification of prescribing outliers. We have applied such an analysis to 6 months of national prescribing data and made the most extreme prescribing outliers in each organization publicly available as interactive dashboards. We intend that these dashboards prompt further qualitative analysis within the individual organizations to identify where service delivery improvements could be made.

Acknowledgments

We are grateful to wider NHS colleagues for discussions that have informed our work on this topic. Conceptualization was done by HC, BM, BG, and AJW. JM, DE, PI, and SB were involved in data curation. LEMH, JM, BM, and AJW performed formal analysis. HC, BG, and AJW were responsible for funding acquisition. LEMH, JM, HC, BM, RC, OM, and AJW conducted investigation. LEMH, JM, HC, BM, RC, and AJW carried methodology. DE, PI, SB, and TOD were involved with the resources. JM, DE, PI, SB, and TOD were responsible for software. Supervision was done by BG. LEMH, JM, HC, BM, and AJW were responsible for visualization. LEMH, JM, HC, BM, and AJW were responsible for writing original draft. LEMH, JM, HC, BM, RC, OM, BG, and AJW were responsible for writing review and editing. BG is the guarantor. This project is funded by the National Institute for Health Research (NIHR) under its Research for Patient Benefit (RfPB) Programme (grant PB-PG-0418-20036). The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care. Funders had no role in the study design, collection, analysis, and interpretation of data; in the writing of the report; and in the decision to submit the paper for publication.

Conflicts of Interest

All authors have completed the International Committee of Medical Journal Editors' uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare the following: BG has received research funding from the Laura and John Arnold Foundation, the NHS National Institute for Health Research (NIHR), the NIHR School of Primary Care Research, the NIHR

Oxford Biomedical Research Centre, the Mohn-Westlake Foundation, NIHR Applied Research Collaboration Oxford and Thames Valley, the Wellcome Trust, the Good Thinking Foundation, Health Data Research UK, the Health Foundation, the World Health Organization, UKRI, Asthma UK, the British Lung Foundation, and the Longitudinal Health and Wellbeing strand of the National Core Studies program; he also receives personal income from speaking and writing for lay audiences on the misuse of science. NJD and JM have been employed on grants received by BG. JM is the recipient of a doctoral studentship from the Wellcome Trust.

Multimedia Appendix 1
The outlier detection dashboard for Devon CCG (including product listings and sparkline plots). The results of our outlier detection methodology are provided as interactive dashboards; here the ten chemicals where prescribing in NHS Devon CCG is higher than most and the ten chemicals where prescribing in NHS Devon CCG is lower than most are presented. Data for each result is highlighted in grey with additional information provided below with no highlighting. BNF Chemical is the chemical of interest (all products represented by this BNF chemical are provided as additional information). Chemical Items provides the number of prescribing items containing this chemical. BNF Subparagraph is the BNF Subparagraph to which the Chemical belongs and Subparagraph Items is the number of prescribing items containing an item belonging to this BNF Subparagraph. Ratio, Mean, std and Z-score place the chemical items count in the context of the subparagraph items count as described in the methods. The sparkline plot provided as additional information for each result shows where the Ratio value for this CCG occurs (vertical red line) in the context of the same Ratio in all CCGs (summarised by the blue line). The y axis is density (see Methods).
[DOCX File , 119 KB - medinform_v11i1e44237_app1.docx]

Multimedia Appendix 2
Prototype dashboard showing the top and bottom five outlying chemicals for Cumbria and northeast STP. BNF Chemical is the chemical of interest, Chemical Items provides the number of prescribing items containing this chemical. BNF Subparagraph is the BNF Subparagraph to which the Chemical belongs and Subparagraph Items is the number of prescribing items containing an item belonging to this BNF Subparagraph. Ratio, Mean, std, and Z_Score place the chemical items count in the context of the subparagraph items count as described in the methods. The sparkline plot shows where the ratio value for this STP occurs (vertical red line) in the context of the same ratio in all STPs (summarised by the blue line).
[DOCX File , 419 KB - medinform_v11i1e44237_app2.docx]

Multimedia Appendix 3
Example dashboard showing the top ten outlying chemicals for Cumbria and northeast STP. BNF Chemical is the chemical of interest, Chemical Items provides the number of prescribing items containing this chemical. BNF Subparagraph is the BNF subparagraph to which the chemical belongs and Subparagraph Items is the number of prescribing items containing an item belonging to this BNF Subparagraph. Ratio, Mean, std, and Z score place the chemical items count in the context of the subparagraph items count as described in the methods. The sparkline plot shows where the ratio value for this STP occurs (vertical red line) in the context of the same Ratio in all STPs (summarised by the blue line). Example dashboard showing the bottom ten outlying chemicals for Cumbria and northeast STP. See Figure S1 for definitions of each column.
[DOCX File , 1142 KB - medinform_v11i1e44237_app3.docx]

References

1. NHS five year forward view. NHS England. URL: https://www.england.nhs.uk/five-year-forward-view/ [accessed 2022-01-26]
2. Getting It Right First Time (GIRFT). URL: https://www.gettingitrightfirsttime.co.uk/ [accessed 2022-01-27]
3. Rightcare methodology. NHS England. URL: https://www.england.nhs.uk/rightcare/what-is-nhs-rightcare/ [accessed 2022-03-15]
4. Information Services Portal (ISP). URL: https://www.nhsbsa.nhs.uk/access-our-data-products/information-services-portal-isp [accessed 2022-02-08]
5. Walker AJ, Curtis HJ, Bacon S, Croker R, Goldacre B. Trends and variation in prescribing of low-priority treatments identified by NHS England: a cross-sectional study and interactive data tool in English primary care. J R Soc Med 2018;111(6):203-213 [FREE Full text] [doi: 10.1177/0141076818769408] [Medline: 29787684]
6. Curtis HJ, Dennis JM, Shields BM, Walker AJ, Bacon S, Hattersley AT, et al. Time trends and geographical variation in prescribing of drugs for diabetes in England from 1998 to 2017. Diabetes Obes Metab 2018;20(9):2159-2168 [FREE Full text] [doi: 10.1111/dom.13346] [Medline: 29732725]
7. Croker R, Walker AJ, Bacon S, Curtis HJ, French L, Goldacre B. New mechanism to identify cost savings in English NHS prescribing: minimising 'price per unit', a cross-sectional study. BMJ Open 2018;8(2):e019643 [FREE Full text] [doi: 10.1136/bmjopen-2017-019643] [Medline: 29439078]
8. Gonem S, Cumella A, Richardson M. Asthma admission rates and patterns of salbutamol and inhaled corticosteroid prescribing in England from 2013 to 2017. Thorax 2019;74(7):705-706. [doi: 10.1136/thoraxjnl-2018-212723] [Medline: 30630892]

9. Saeed HS, Wright RB, Ghosh SK. Trends in the prescribing of topical nasal agents using an NHS England data base. Clin Otolaryngol 2018;43(5):1296-1302. [doi: 10.1111/coa.13143] [Medline: 29770588]
10. Ivers N, Jamtvedt G, Flottorp S, Young JM, Odgaard-Jensen J, French S. Audit and feedbackffects on professional practice and healthcare outcomes. Cochrane Database Syst Rev 2012:CD000259. [doi: 10.1002/14651858.cd000259.pub3]
11. Rodgers S, Avery AJ, Meechan D, Briant S, Geraghty M, Doran K, et al. Controlled trial of pharmacist intervention in general practice: the effect on prescribing costs. Br J Gen Pract 1999;49(446):717-720 [FREE Full text] [Medline: 10756613]
12. ePACT2. NHS Business Services Authority. URL: https://www.nhsbsa.nhs.uk/access-our-data-products/epact2 [accessed 2023-02-08]
13. OpenPrescribing. URL: https://openprescribing.net/ [accessed 2022-09-22]
14. Hirsch O, Donner-Banzhoff N, Schulz M, Erhart M. Detecting and visualizing outliers in provider profiling using funnel plots and mixed effects models-an example from prescription claims data. Int J Environ Res Public Health 2018;15(9):2015 [FREE Full text] [doi: 10.3390/ijerph15092015] [Medline: 30223551]
15. MacKenna B, Curtis HJ, Hopcroft LEM, Walker AJ, Croker R, Macdonald O, et al. Identifying patterns of clinical interest in clinicians' treatment preferences: hypothesis-free data science approach to prioritizing prescribing outliers for clinical review. JMIR Med Inform 2022;10(12):e41200 [FREE Full text] [doi: 10.2196/41200] [Medline: 36538350]
16. Yu Y, Wilson M, King CE, Hill R. Up-scheduling and codeine supply in Australia: analysing the intervention and outliers. Addiction 2021;116(12):3463-3472. [doi: 10.1111/add.15566] [Medline: 33999465]
17. Walker AJ, Bacon S, Croker R, Goldacre B. Detecting change in comparison to peers in NHS prescribing data: a novel application of cumulative sum methodology. BMC Med Inform Decis Mak 2018;18(1):62 [FREE Full text] [doi: 10.1186/s12911-018-0642-6] [Medline: 29986693]
18. Primary care networks. NHS England. URL: https://www.england.nhs.uk/primary-care/primary-care-networks/ [accessed 2022-02-25]
19. Clinical commissioning groups. NHS England. URL: https://www.england.nhs.uk/ccgs/ [accessed 2022-02-25]
20. Sustainability and transformation plans (STPs) explained. The King's Fund. 2017. URL: https://www.kingsfund.org.uk/topics/integrated-care/sustainability-transformation-plans-explained [accessed 2022-02-25]
21. Angus L. The role and functions of CCG medicines optimisation teams. NHS Clinical Commissioners. 2021. URL: https://www.nhsconfed.org/system/files/2021-07/Role-and-functions-of-the-CCG-medicines-optimisation-team.pdf [accessed 2023-03-17]
22. English prescribing data (EPD). NHS Business Services Authority. URL: https://www.nhsbsa.nhs.uk/prescription-data/prescribing-data/english-prescribing-data-epd [accessed 2023-02-08]
23. Curtis HJ, Goldacre B. OpenPrescribing: normalised data and software tool to research trends in English NHS primary care prescribing 1998-2016. BMJ Open 2018;8(2):e019921 [FREE Full text] [doi: 10.1136/bmjopen-2017-019921] [Medline: 29476029]
24. GP and GP practice related data. NHS Digital. URL: https://digital.nhs.uk/services/organisation-data-service/file-downloads/gp-and-gp-practice-related-data [accessed 2022-03-16]
25. OpenPrescribing Outlier Detection. URL: https://openprescribing.net/labs/outlier_reports/ [accessed 2022-08-10]
26. Scott DW. Multivariate Density Estimation: Theory, Practice, and Visualization. New York: Wiley; 1992.
27. Outlier detection dashboards: code for data management and analysis. URL: https://github.com/ebmdatalab/openprescribing/blob/main/openprescribing/pipeline/management/commands/outlier_reports.py [accessed 2022-09-22]
28. 12.1.1 Otitis externa. North & East Devon Formulary and Referral. URL: https://northeast.devonformularyguidance.nhs.uk/formulary/chapters/12.-ear-nose-oropharynx/12.1-ear/12-1-1-otitis-externa [accessed 2022-03-29]
29. 5.1.12 Quinolones. North & East Devon Formulary and Referral. URL: https://southwest.devonformularyguidance.nhs.uk/formulary/chapters/5.-infections/5.1-antibacterial-drugs/quinolones [accessed 2022-03-29]
30. 11.3.1 Antibacterials. North & East Devon Formulary and Referral. URL: https://northeast.devonformularyguidance.nhs.uk/formulary/chapters/11.-eye/11.3-eye-infections/11-3-1-antibacterials [accessed 2022-03-29]
31. Drug tariff for Tariff prices for Fusidic acid 1% modified-release eye drops. URL: https://openprescribing.net/tariff/?codes=1103010H0AAAAAA [accessed 2022-09-22]
32. Management of constipation in adults. North & East Devon Formulary and Referral. URL: https://northeast.devonformularyguidance.nhs.uk/formulary/chapters/1.-gastrointestinal/constipation [accessed 2022-03-29]
33. Topical corticosteroids. North & East Devon Formulary and Referral. URL: https://northeast.devonformularyguidance.nhs.uk/formulary/chapters/13.-skin/13-4-corticosteroids-topical [accessed 2022-03-29]
34. Walker AJ, Curtis HJ, Bacon S, Croker R, Goldacre B. Trends, geographical variation and factors associated with prescribing of gluten-free foods in English primary care: a cross-sectional study. BMJ Open 2018;8(3):e021312 [FREE Full text] [doi: 10.1136/bmjopen-2017-021312] [Medline: 29661914]
35. Prescribing gluten-free foods in primary care: guidance for CCGs. NHS England. 2018. URL: https://www.england.nhs.uk/publication/prescribing-gluten-free-foods-in-primary-care-guidance-for-ccgs/ [accessed 2023-02-21]
36. Curtis HJ, Croker R, Walker AJ, Richards GC, Quinlan J, Goldacre B. Opioid prescribing trends and geographical variation in England, 1998–2018: a retrospective database study. Lancet Psychiatry 2019;6(2):140-150. [doi: 10.1016/s2215-0366(18)30471-1]

37. Data saves lives: reshaping health and social care with data (draft). URL: https://www.gov.uk/government/publications/data-saves-lives-reshaping-health-and-social-care-with-data-draft/data-saves-lives-reshaping-health-and-social-care-with-data-draft [accessed 2021-06-30]
38. Goldacre B, Morley J. Better, broader, safer: using health data for research and analysis. UK Government Department of Health and Social Care. 2022 Apr 07. URL: https://www.gov.uk/government/publications/better-broader-safer-using-health-data-for-research-and-analysis [accessed 2023-02-21]

Abbreviations

BNF: British National Formulary
CCG: clinical commissioning group
MO: medicines optimization
NHS: National Health Service
PCN: primary care network
STP: sustainability and transformation partnership

Edited by A Benis; submitted 11.11.22; peer-reviewed by C Tolley, H De Loof, L Monteiro, S Leitch; comments to author 05.01.23; revised version received 24.02.23; accepted 11.03.23; published 19.04.23.

Please cite as:
Hopcroft LEM, Massey J, Curtis HJ, Mackenna B, Croker R, Brown AD, O'Dwyer T, Macdonald O, Evans D, Inglesby P, Bacon SCJ, Goldacre B, Walker AJ
Data-Driven Identification of Unusual Prescribing Behavior: Analysis and Use of an Interactive Data Tool Using 6 Months of Primary Care Data From 6500 Practices in England
JMIR Med Inform 2023;11:e44237
URL: https://medinform.jmir.org/2023/1/e44237
doi:10.2196/44237
PMID:37074763

Predicting Postoperative Hospital Stays Using Nursing Narratives and the Reverse Time Attention (RETAIN) Model: Retrospective Cohort Study

Sungjoo Han[1], BS; Yong Bum Kim[2], MD, PhD; Jae Hong No[2], MD, PhD; Dong Hoon Suh[2], MD, PhD; Kidong Kim[2], MD, PhD; Soyeon Ahn[1], PhD

[1]

[2]

Corresponding Author:
Soyeon Ahn, PhD

Abstract

Background: Nursing narratives are an intriguing feature in the prediction of short-term clinical outcomes. However, it is unclear which nursing narratives significantly impact the prediction of postoperative length of stay (LOS) in deep learning models.

Objective: Therefore, we applied the Reverse Time Attention (RETAIN) model to predict LOS, entering nursing narratives as the main input.

Methods: A total of 354 patients who underwent ovarian cancer surgery at the Seoul National University Bundang Hospital from 2014 to 2020 were retrospectively enrolled. Nursing narratives collected within 3 postoperative days were used to predict prolonged LOS (≥10 days). The physician's assessment was conducted based on a retrospective review of the physician's note within the same period of the data model used.

Results: The model performed better than the physician's assessment (area under the receiver operating curve of 0.81 vs 0.58; *P*=.02). Nursing narratives entered on the first day were the most influential predictors in prolonged LOS. The likelihood of prolonged LOS increased if the physician had to check the patient often and if the patient received intravenous fluids or intravenous patient-controlled analgesia late.

Conclusions: The use of the RETAIN model on nursing narratives predicted postoperative LOS effectively for patients who underwent ovarian cancer surgery. These findings suggest that accurate and interpretable deep learning information obtained shortly after surgery may accurately predict prolonged LOS.

(JMIR Med Inform 2023;11:e45377) doi:10.2196/45377

KEYWORDS

discharge prediction; text mining; free text; extraction; length of stay; hospital stay; electronic health record; EHR; discharge; interpretable deep learning; risk prediction; nursing; machine learning; deep learning; predict; ovarian cancer

Introduction

Postoperative length of stay (LOS) is an important indicator of hospital management efficiency. A precise estimate of LOS optimizes hospital bed availability and resource allocation, thereby improving health outcomes and lowering costs [1,2]. There is an increasing need to predict LOS using electronic health records (EHRs) with machine learning methods [3-6]. EHRs contain data on patients' demographics, diagnoses, medications, vital signs, and laboratory results, which are fed into deep learning algorithms. For example, Safavi et al [7] have suggested a feedforward neural network model comprising clinical and administrative data extracted from EHRs to predict discharge from inpatient surgical care. Zhang et al [8] have investigated a prediction model for next-day discharge using EHR access logs combined with gradient-boosted ensembles of decision trees. For this study, we refer to Stone et al [9] for a comprehensive review of the prediction of hospital LOS.

We focused on nursing narratives in EHRs as a promising predictor of postoperative LOS. Nursing narratives are representations of the nursing process and contain data regarding when and how nursing actions are performed on patients [10,11]. Analyses of nursing notes using machine learning models have shown promising results in predicting short-term patient outcomes [12,13]. We have previously reported that a deep learning model based on nursing narratives can effectively predict postoperative LOS [14]. However, a fundamental problem of deep learning models is their lack of interpretability, which restrains their clinical applicability [15,16]. Moreover, our previous study implemented long short-term memory using frequencies of individual nursing narrative entries for 5

postoperative days as features, which limited the power of dependencies between time steps of sequence data.

To overcome this issue, various interpretable artificial intelligence models have been examined [17]. The Reverse Time Attention (RETAIN) model is an interpretable predictive model developed for application with EHR data. RETAIN's major advantage is its high accuracy while remaining clinically interpretable by adapting a 2-level neural attention mechanism in a recurrent neural network architecture. Consequently, RETAIN can detect both influential nurse visits and clinical features [16]. Several studies have demonstrated the clinical utility of the RETAIN model in diverse clinical contexts [18-23]. AlSaad et al [21] have shown a simplified version of the RETAIN architecture that significantly predicted preterm birth and enabled individual-level prediction explanations at the visitation level and medical code level (*International Classification of Diseases, Ninth Revision* [*ICD-9*] or *ICD-10* codes). Rasmy et al [23] have adapted a language model that combined the RETAIN model with two independent EHR databases; this model achieved a high degree of accuracy in predicting both heart failure and the onset of pancreatic cancer.

In this study, we examined the performance of an interpretable deep learning model using longitudinal nursing narratives to predict prolonged LOS and extracted the significant nursing narrative features to better understand the prediction model.

Methods

Ethical Considerations

This study was approved by the Seoul National University Bundang Hospital (SNUBH) Institutional Review Board (B-2011/646-104 for model development; B-2103-675-101 for physician comparison).

Setting

ICD-10 diagnosis code C56 was used to identify the study population. Data were retrospectively collected from patients admitted to the SNUBH for first-time ovarian cancer surgery between 2014 and 2021.

We divided the data into two parts by period: the internal data set (collected between January 2014 and September 2020) and the external validation data set (collected between October 2020 and February 2021) [24]. We chose the most recent 5 months of data as the external validation data set. The internal data set was used for training, validating, and testing the model, while the external data set was used for evaluating the final performance of the model and comparing the results with the physician's assessment.

The exclusion criteria included readmission, admission with postoperative LOS <3 days, and patients who underwent surgery <20 times in the internal data set (Figure S1 in Multimedia Appendix 1). Postoperative LOS was chosen as the outcome variable because in-hospital LOS can be affected by several nonoperational factors such as patient characteristics or social circumstances, type of admission, patient place of residence, emergencies, or weekend admissions [9,25,26]. The postoperative LOS was defined as the number of days from the date of the index operation to the date of discharge, where the date of the index operation, denoted as day 0, was defined as the date in any of the operation-related nursing narratives. For example, if a patient was discharged on day 0, the postoperative LOS was 1.

Nursing Narratives

We extracted nursing narratives chronologically. At SNUBH, nursing narratives are easily integrated into a structured database because individual features are mapped to unique codes [11,27]. For instance, the nursing narrative "checked the vital signs" is mapped to code N1, whereas "no dizziness" is mapped to code N2. A nurse enters patient statuses in the EHR system by searching nursing narratives using keywords such as "vital" or "dizzy" and selecting the appropriate nursing narratives from the list of related narratives. Some nursing narratives allow for the inclusion of additional information such as body temperature or free text [14]. Consequently, patient information was entered as a combination of unique codes (eg, N1, checked the vital signs, or N2, checked whether the patient felt dizzy), code entry time, and a specific value such as body temperature. These structured nursing narrative sets allowed us to retrieve patient information without the need for natural language preprocessing.

RETAIN Architecture

Prolonged LOS was defined as events ≥10 postoperative days, which was the third quantile of postoperative LOS in both the internal and external validation data sets. Our results showed that the volume of nursing narratives entered within 3 postoperative days was high and tended to decrease afterward; therefore, we decided to use patient information within 3 postoperative days, that is, from day 0 to day 2 (Figure S2 in Multimedia Appendix 1). The extracted time series of nursing narratives and corresponding unique codes were inverted to 3D arrays (patients, postoperative days, and nursing narratives' unique codes).

The internal data set was randomly split, allocating 60% (n=192) of participants to the training set and 20% (n=64) of participants each to the validation and testing sets. The training set was used to train the models, the validation set was used to determine the values of the hyperparameters that increase the area under the receiver operating curve (AUC), and the test set was used to evaluate the performance of the best model. The best model was also applied to the external validation data set. Therefore, the performance of the best model was measured twice (the test set of the internal data set and the external validation data set). Furthermore, we compared the performance of the external validation data set with the physician's assessment. The RETAIN model was constructed with two neural attentions that can identify influential nurse visits and meaningful features. The RETAIN model uses linear embedding to enhance interpretability. The contribution score was calculated using visit-level attention weights, variable-level attention weights, and embedding weights.

The default settings for RETAIN were used. L2 regularization for the final classifier weight, input embedding weight, and alpha-generating weight was set to 0.0001 for all models. Following a learning process using batch sizes of 8, 16, and 32,

the model with the highest AUC performance on the test set was selected as the best model. If models had the same AUC value, the one with the highest sensitivity was selected as the best model.

Model Interpretation and Influential Features Extraction

The RETAIN model reported the contribution scores that represented the extent to which each feature contributed to the prediction. In this study, features with a high contribution score were associated with a high likelihood of prolonged LOS. We identified the input features with high contribution scores as influential features, which showed a significant difference between prolonged and short LOS via a *t* test with a *P* value cutoff of .05.

Comparison Between the Deep Learning Model and a Physician's Expert Clinical Assessment

We compared the deep learning model and a physician's assessment vis-à-vis their predictive capability for prolonged LOS. A gynecologic oncologist with 15 years of experience reviewed patients' demographics, progress notes, surgical reports, and clinical notes available within 3 postoperative days. Blinded to the final discharge date, the physician predicted whether patients would experience prolonged LOS. The DeLong test was used to compare the AUCs of the deep learning model and a physician assessment [28].

Visualization and Statistical Analysis

Statistical analyses were performed using Python (version 3.9.13; Python Software Foundation), RETAIN (version 1.0; Edward Choi) [16], and R (version 4.0.5; R Foundation for Statistical Computer) software. The influential features were visualized using the ComplexHeatmap package in R software.

Results

Patient Characteristics

This study retrospectively enrolled 354 patients (n=320 in the internal data set and n=34 in the external validation data set; mean age 54, SD 13 years; Table S1 and Figure S1 in Multimedia Appendix 1). A total of 51,603 nursing narratives in the internal data set composed the model inputs. Patients in the prolonged LOS group were older (for instance, in the internal data set, the mean age was 57, SD 12 years and 52, SD 14 years in the prolonged LOS and short LOS groups, respectively; *P*=.002) and had higher total nursing narrative volumes (mean 188, SD 62 vs mean 150, SD 33 narratives within 3 postoperative days; *P*<.001). The. nursing narrative entries per nurse visit were similar (mean 5.6, SD 8.4 vs mean 5.9, SD 7.9 for the prolonged LOS and short LOS groups, respectively; *P*=.64), but more frequent nurse visits were observed in the prolonged LOS group (mean 33, SD 9 vs mean 25, SD 9 visits within 3 postoperative days; *P*=.03).

Prediction of Prolonged LOS via RETAIN

The experimental scheme is shown in Figures S2 and S3 in Multimedia Appendix 1. The RETAIN model was developed using the internal data set, while the model's performance was calculated and compared to a physician's expert clinical assessment using the external validation data set. The predictive contribution score derived from the RETAIN model indicates the probability of prolonged LOS. Patients with a final predictive score >0.5 were classified as expected prolonged LOS events. The RETAIN model reported the scores for each patient and nursing narrative; these scores were used to determine highly influential nursing narratives. We determined potent nursing narratives for each patient based on patient-wise normalization scores, after applying normalization using a patient-centric mean and SD. Thereafter, influential nursing narratives were defined as those consistently showing a statistically significant difference in the raw contribution score between the prolonged and short LOS groups.

Table 1 shows a comparison between the model's performance and a physician's expert clinical assessment, considering various nursing narrative sets. The model trained with nursing narratives showed an AUC of 0.81. The deep learning model performed better than the physician's assessment (AUC 0.58; *P*=.02; Figure S4 in Multimedia Appendix 1).

Table . Model performance on the internal and external validation set.

Data set and model	AUC[a]	Accuracy	Sensitivity	Specificity	F_1-score	P value[b]
Internal data set						
RETAIN[c] with nursing narratives	0.76	0.80	0.55	0.91	0.63	N/A[d]
External validation data set [e]						.02
RETAIN with nursing narratives	0.81	0.85	0.44	1.00	0.62	
Physician assessment	0.58	0.65	0.44	0.72	0.40	

[a]AUC: area under the receiver operating curve.

[b]The DeLong test was conducted to compare the AUCs of the RETAIN model and physician assessment.

[c]RETAIN: Reverse Time Attention.

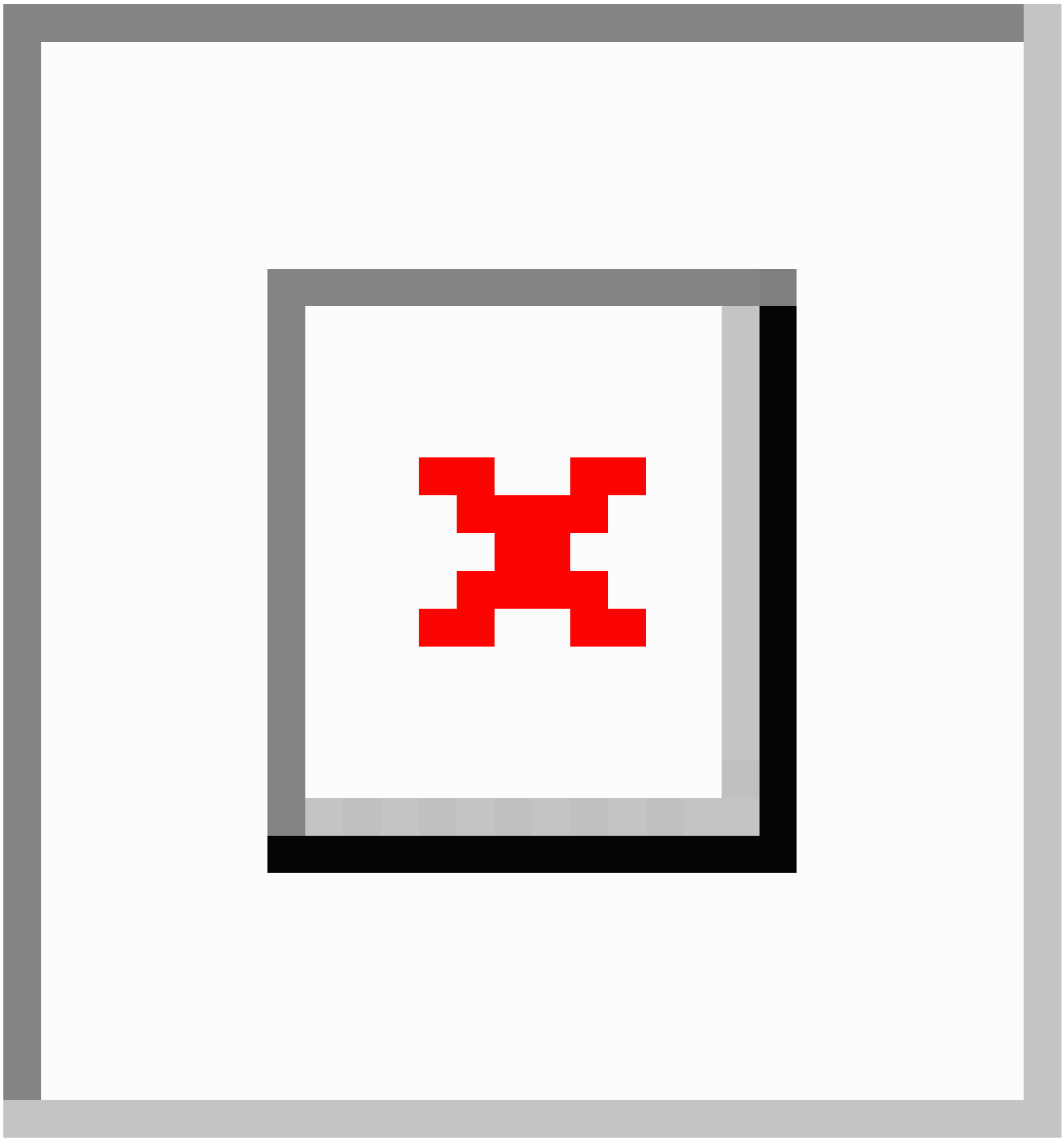
[d]N/A: not applicable.

[e]The RETAIN model performance and physician assessment were compared using the external validation data set. A total of 34 patients were available for 3 postoperative days.

Influence of Nurse Visits on the Prediction of Prolonged LOS

Examples of contribution score graphs were visualized for patients in the prolonged and short LOS groups (Figure 1). As expected, the patients in the prolonged LOS group exhibited high contribution scores. Nurse visits on the first postoperative day (ie, day 1) were identified as highly influential because nursing narratives entered on that day exhibited higher contribution scores.

Figure 1. Highly influential nursing narrative (NN) examples presenting the differences between the prolonged and short LOS groups' contribution score graphs. NNs are arranged in chronological order, while the corresponding scores are represented as dots. The predictive score indicates the probability of prolonged LOS, which was estimated by the Reverse Time Attention model, with (A) a postoperative LOS of 11 days (predictive score: 0.90) and (B) a postoperative LOS of 4 days (predictive score: 0.01). LOS: length of stay.



Highly influential narratives showing statistically significant differences in contribution scores between the prolonged and short LOS groups included the following: "confirmed by a doctor," "injected intravenous patient-controlled analgesia [PCA]," "injected intravenous fluids," "no PCA side effects," "observed the pattern of Jackson-Pratt [J-P] tube drainage," "patient's pain in surgical area was tolerable," "provided mental support," "maintained J-P tube," "maintained Foley catheter," "no oozing in the drainage tube insertion area," "measured body temperature," "provided safety care," and "notified a doctor" (Figures 2 and 3). The three most influential narratives (according to their lower *P* values) were "confirmed by a doctor," "injected intravenous PCA," and "injected intravenous fluids" (Table 2), whose contribution scores were visualized by *t*-distributed stochastic neighbor embedding (Figure 4).

Figure 2. Heat map visualizing the contribution scores of highly influential nursing narratives (NNs). NN-level normalized contribution scores were calculated for patients of the external data set. The *P* value represents the results of the *t* test for raw contribution score comparison between the prolonged LOS and short LOS groups. J-P: Jackson-Pratt; LOS: length of stay; PCA: patient-controlled analgesia.

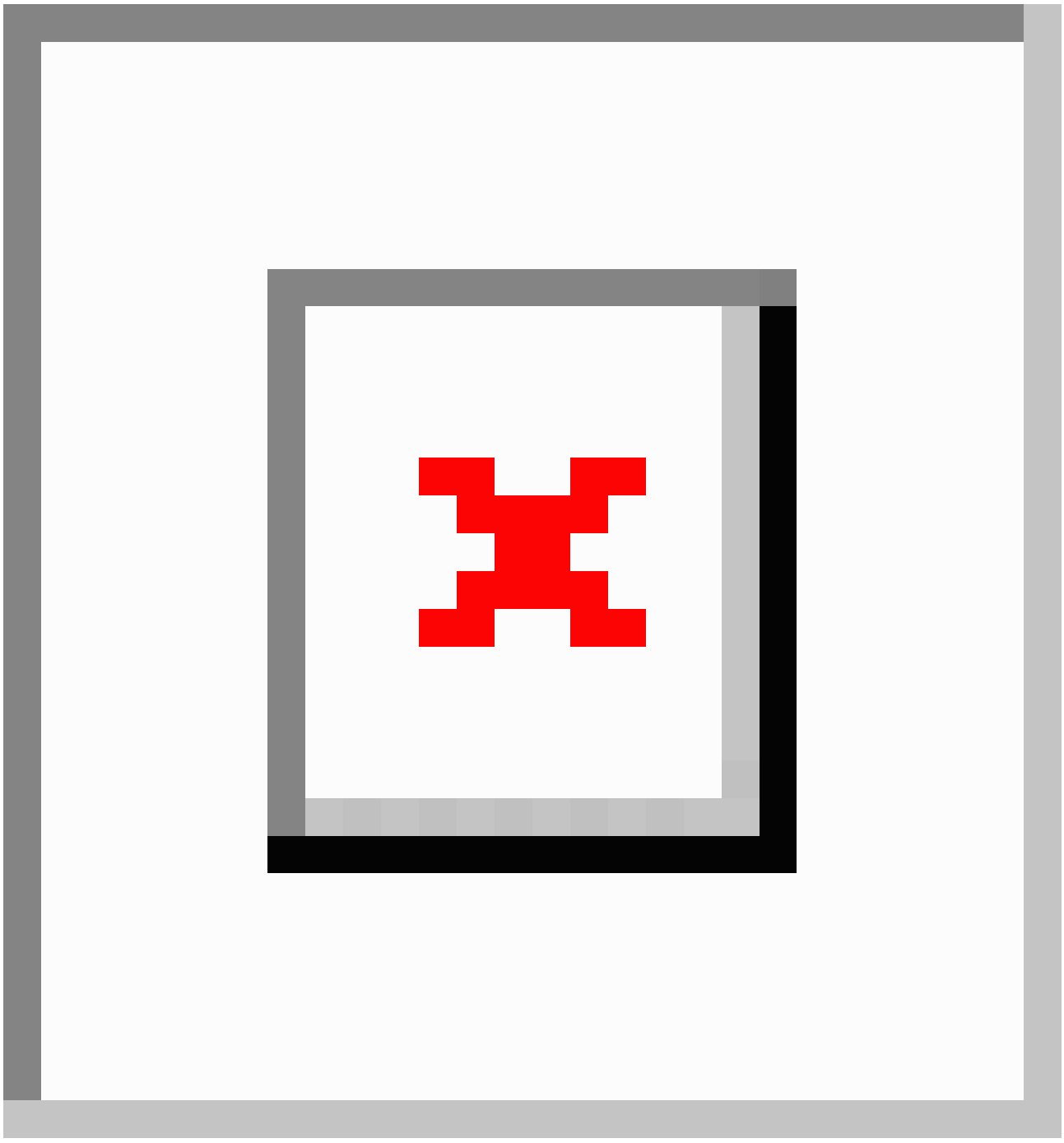


Figure 3. The contribution score graph highlights highly influential nursing narratives (NNs) of the prolonged LOS group, with the NNs arranged in chronological order. The areas of the corresponding contribution scores are filled. The predictive score indicates the probability of a prolonged LOS. The patients with predictive scores >0.5 were classified as expected prolonged LOS. The most influential NNs are represented as orange dots. (A) Postoperative LOS: 11 days; predictive score: 0.9; (B) postoperative LOS: 10 days, predictive score: 0.67; (C) postoperative LOS: 14 days, predictive score: 0.63. J-P: Jackson-Pratt; LOS: length of stay; PCA: patient-controlled analgesia.

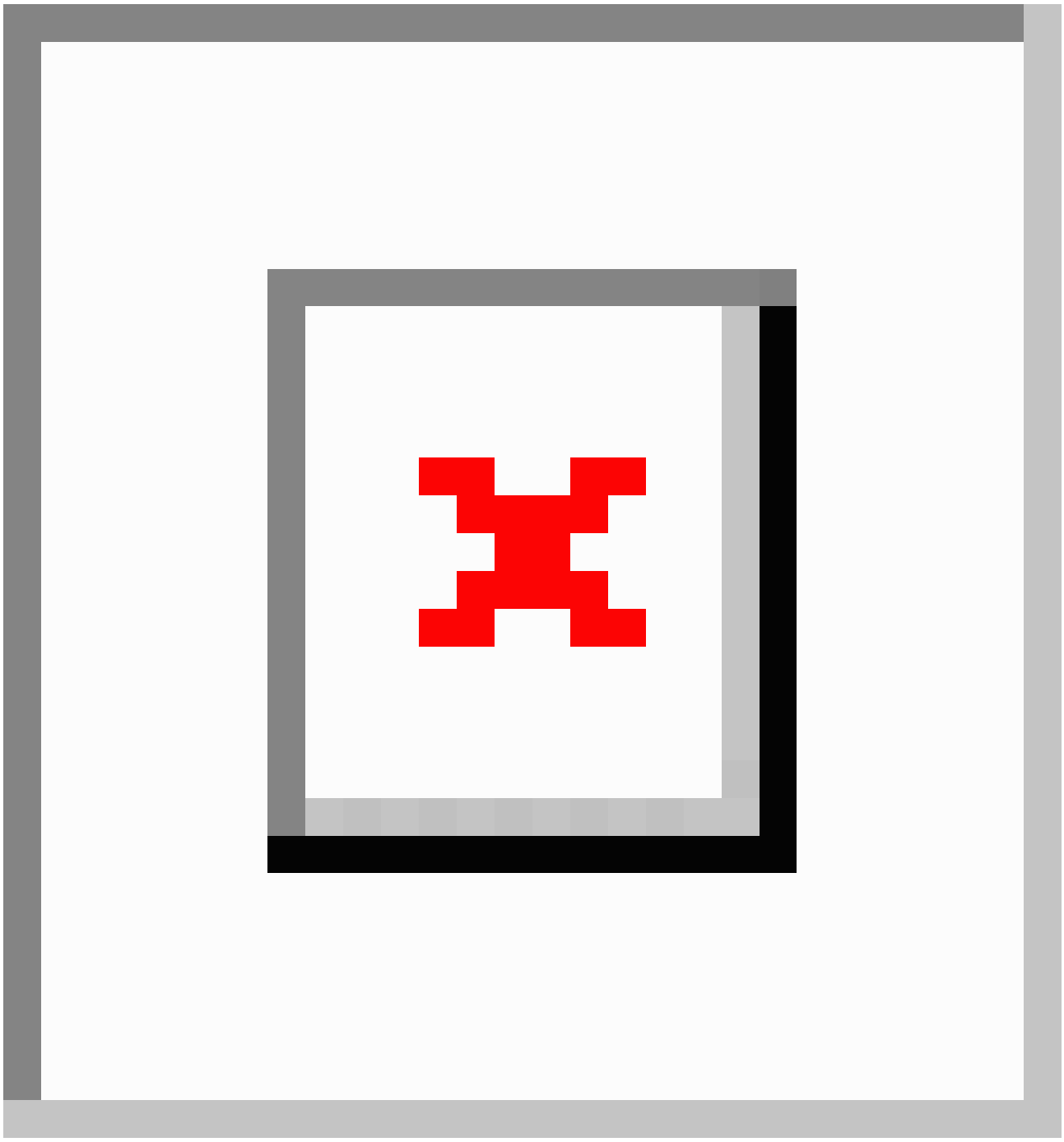


Table . Contribution scores of influential nursing narratives.

Nursing narratives	Internal data set (n=320)			External validation data set (n=34)		
	Prolonged LOS[a], mean (SD)	Short LOS, mean (SD)	*P* value[b]	Prolonged LOS, mean (SD)	Short LOS, mean (SD)	*P* value[b]
Confirmed by a doctor	0.044 (0.038)	–0.001 (0.021)	<.001	0.030 (0.033)	0.000 (0.020)	.002
Injected intravenous PCA[c]	0.012 (0.046)	–0.074 (0.041)	<.001	–0.030 (0.047)	–0.076 (0.035)	.003
Injected intravenous fluids	0.012 (0.044)	–0.066 (0.036)	<.001	–0.009 (0.043)	–0.057 (0.036)	.003
No PCA side effects	–0.002 (0.058)	–0.092 (0.052)	<.001	–0.037 (0.055)	–0.094 (0.044)	.004
Observed the pattern of J-P[d] tube drainage	0.016 (0.039)	–0.039 (0.033)	<.001	–0.001 (0.038)	–0.039 (0.034)	.007
Patient's pain in the surgical area was tolerable	0.019 (0.036)	–0.031 (0.027)	<.001	–0.011 (0.025)	–0.034 (0.024)	.02
Provided mental support	–0.004 (0.043)	–0.065 (0.060)	<.001	–0.009 (0.026)	–0.053 (0.056)	.03
Maintained J-P tube	0.019 (0.036)	–0.035 (0.031)	<.001	–0.005 (0.040)	–0.034 (0.031)	.03
Maintained Foley catheter	0.019 (0.019)	0.002 (0.007)	<.001	0.011 (0.017)	0.003 (0.006)	.045
No oozing in the drainage tube insertion area	0.011 (0.022)	–0.008 (0.017)	<.001	0.001 (0.004)	–0.009 (0.016)	.06
Measured body temperature	0.008 (0.025)	–0.018 (0.027)	<.001	0.003 (0.009)	–0.007 (0.017)	.11
Provided safety care	0.026 (0.027)	0.006 (0.012)	<.001	0.020 (0.019)	0.011 (0.016)	.18
Notified a doctor	0.021 (0.019)	0.004 (0.010)	<.001	0.012 (0.017)	0.006 (0.010)	.24

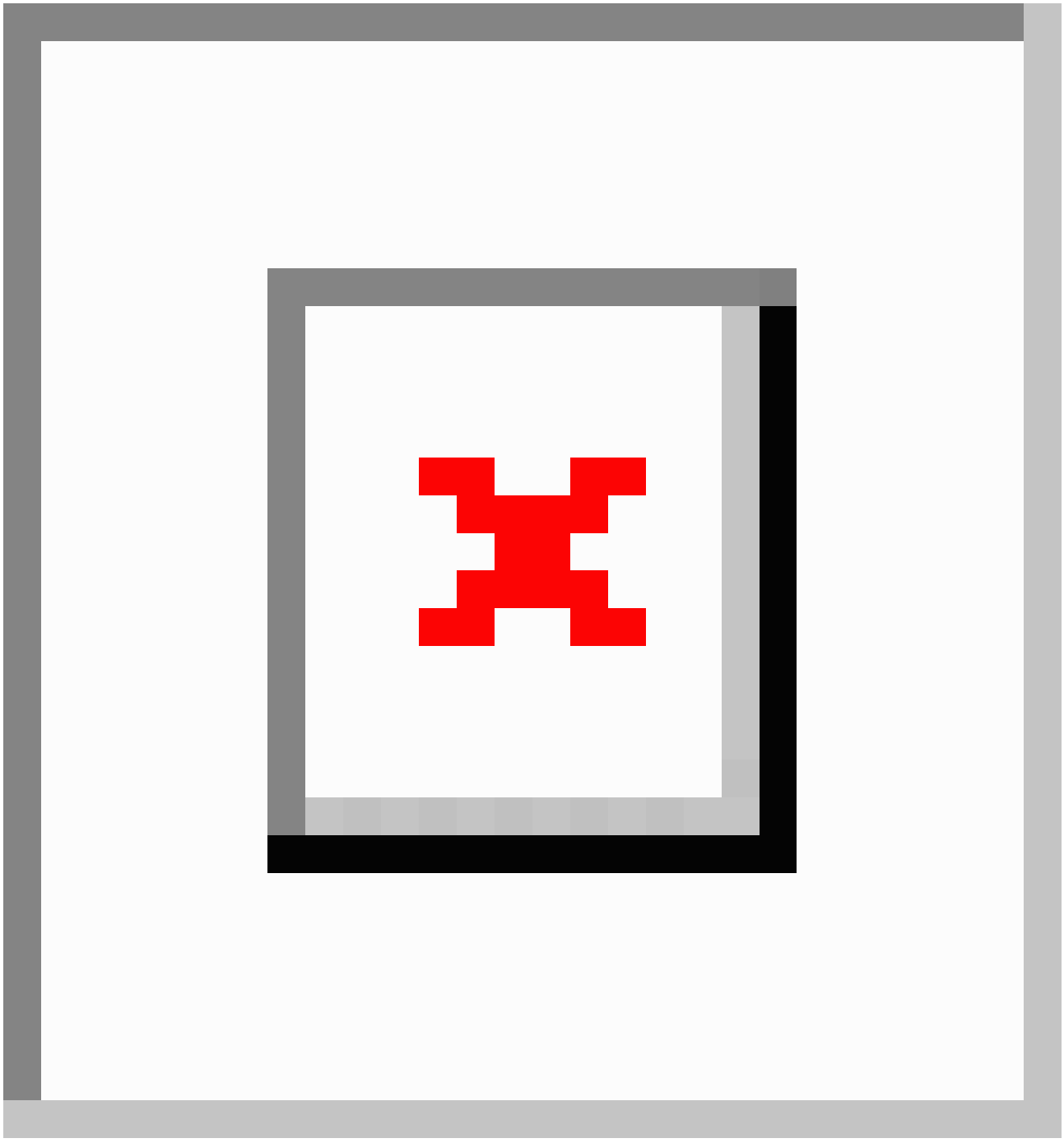
[a]LOS: length of stay.

[b]*P* values represent the results of the *t* test for raw contribution scores compared between the prolonged and short LOS groups.

[c]PCA: patient-controlled analgesia.

[d]J-P: Jackson-Pratt.

Figure 4. T-distributed stochastic neighbor embedding plot using the top three highly influential nursing narratives; the contribution score was estimated from the external data set. LOS: length of stay; PCA: patient-controlled analgesia.



Once the three most influential nursing narratives were identified, we further investigated the total number of entries and the first entry time since the handoff. The "confirmed by a doctor" narrative reoccurred in the prolonged LOS group (mean 5.8, SD 4.4 vs mean 3.1, SD 2.3 nursing narratives in the prolonged and short LOS groups, respectively) and was entered earlier (Table S2 in Multimedia Appendix 1). Conversely, the narratives "injected intravenous PCA" and "injected intravenous fluids" exhibited similar entry values but were entered a few hours later in the prolonged LOS group.

Discussion

Principal Results

In this study, a RETAIN model was used to predict postoperative LOS using nursing narratives. The model achieved a higher AUC value of 0.81 compared to the physician assessment's AUC of 0.58 (P=.02). Highly influential nursing narratives were identified that differed in their contribution scores between the prolonged and short LOS groups, including confirming by a doctor, administering intravenous PCA, and providing intravenous fluids.

To our knowledge, this is the first study to extract nursing narratives' influential features by normalizing contribution scores estimated via RETAIN. By investigating these influential features, we discovered that the volume and timing of individual narratives are key factors. The likelihood of prolonged LOS increases if a physician must check the patient more often or if intravenous fluids or intravenous PCA are administered late. This degree of interpretability was not achievable in previous studies that relied on volume-centric statistical methods and conventional deep learning models.

Strengths

This study demonstrated that nursing narratives can accurately predict the postoperative LOS of patients who underwent surgery for ovarian cancer. We implemented an interpretable deep learning model to identify highly influential nursing narratives. Notably, nursing narratives entered one day after surgery were the primary predictors for prolonged LOS.

Nursing narratives, serving as proxies for the care given to patients, demonstrated predictive value for LOS. Nursing narratives thus reflect the actions and interventions carried out by health care professionals. By identifying highly influential nursing narratives and presenting the different action timing and volume of each narrative, we enhanced the model's interpretability and showed that the relevant nursing activities could serve as indicators for LOS.

These findings support other studies that have shown that nursing notes may predict short-term patient outcomes more accurately than physician notes [29,30]. Nurses frequently summarize patients' situations by describing their symptoms, as well as their nursing actions and responses, without the restriction of structured forms [31-33]. Thus, nursing notes serve as a snapshot of patients' current statuses and exhibit a higher degree of freedom compared to physicians' notes, which provide a problem-focused summary. In a prospective cohort of patients who are critically ill, nurses predicted in-hospital mortality slightly more accurately than physicians, whereas the latter predicted long-term outcomes more accurately [29]. Huang et al [30] applied natural language processing to free-text nursing notes to predict multiple outcomes, including prolonged hospital stay or mortality, using the Multiparameter Intelligent Monitoring of Intensive Care III. This study also acknowledged the superior predictability value of nursing notes over physicians' notes when using refined features within the first 48 hours of admission. However, none of these studies presented the additional interpretation of specific nursing notes.

Furthermore, this study showed that the total volume of nursing narratives is a significant factor for prolonged LOS, which was consistent with previous studies conducted in different settings. Schnock et al [34] have conducted a multicenter qualitative study in intensive and acute care units to discover nursing documentation patterns indicating recovery patterns. Woo et al [35] have used the natural language processing of nursing notes from patients admitted to home care and found that the frequency of wound infection–related text in nursing notes increased before hospitalization or emergency department visits. However, these studies faced a common barrier to using nursing notes: the extraction of standardized information. Accordingly, there is a significant need for health care providers to standardize nursing assessments and free-text notes [30].

We showed that the nursing narratives "confirmed by a doctor," "injected intravenous PCA," and "injected intravenous fluids" were relevant to a prolonged stay for patients with surgical procedures. These narratives suggested that a patient's condition is complicated, and additional support for pain management or fluid management was required. Timely communication and collaboration between nursing and medical staff, effective pain management, and appropriate fluid management are important considerations in surgical patient care, which can impact the LOS and overall patient outcomes.

Limitations

This study had several limitations. First, this study was based on data from a single-hospital EHR system. The EHR system at SNUBH allows for the standardization of nursing narratives, which enables the creation of a structured database. However, in most hospitals, free-text nursing notes are common; therefore, the preprocessing of natural language is required to generalize this study's findings. As a starting point, it is worthwhile to examine the highly influential nursing narratives identified in our study. Second, we chose nursing narratives entered within a 3-day postoperative interval, which can be shortened in future studies. For example, 2-day postoperative data have been used in several studies [9,30]. Furthermore, a strategic patient care plan that combines a short-interval model and a long-interval model could be developed. Third, this study's sample size was small, while the model was developed in a single-disease setting. To consider the dependency of nursing narratives according to different surgery and patient settings, transfer learning (in which a model trained in a larger population is fine-tuned with an independent surgery setting) can be considered. Future studies with multiple hospital settings and multimodal features are required [36]. Fourth, like other machine learning models, training and testing a model requires a large amount of data, and multiple validation sets are needed to avoid overfitting [37]. In addition, as the RETAIN model receives input values for each variable and visit level, it may be difficult to apply the model to unstructured data such as free text or data that cannot be classified by date. Finally, it is important to acknowledge that physician assessments were done retrospectively, potentially not capturing dynamic clinical situations.

Future Perspectives

Collecting a larger data set that includes a wider range of patients and additional predictors such as laboratory data or comorbidity information is essential. We firmly believe that integrating nursing narratives with broader information, including physician assessments, can lead to a better prediction model.

Conclusions

In this study, an interpretable prediction model for a longer postoperative LOS was developed using nursing narratives. The day after surgery was the most critical time for prediction, and influential nursing narratives were revealed. Although nursing narratives serve as proxies for the care given to patients, our study suggests that they have the potential to be predictors for

LOS. The developed model can help identify patients with a prolonged hospital stay at the right time, thereby improving patient care and reducing hospital management burden. To strengthen the evidence supporting the predictive value of nursing narratives, either alone or in combination with broader information such as physician assessment, a larger data set would be beneficial.

Our study highlights that nursing narratives are predictors for prolonged LOS in patients undergoing ovarian cancer surgery. We emphasize the comprehensive nature of nursing actions and their timing in predicting patient outcomes and suggest methods to incorporate into a prediction model.

Acknowledgments

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea, which is funded by the Ministry of Education (NRF-2020R1C1C1007704 to SA).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary material.
[DOCX File, 2122 KB - medinform_v11i1e45377_app1.docx]

References

1. Gonçalves-Bradley DC, Lannin NA, Clemson LM, Cameron ID, Shepperd S. Discharge planning from hospital. Cochrane Database Syst Rev 2016 Jan 27;2016(1):CD000313. [doi: 10.1002/14651858.CD000313.pub5] [Medline: 26816297]
2. Parikh RB, Kakad M, Bates DW. Integrating predictive analytics into high-value care: the dawn of precision delivery. JAMA 2016 Feb 16;315(7):651-652. [doi: 10.1001/jama.2015.19417] [Medline: 26881365]
3. Bacchi S, Gluck S, Tan Y, et al. Prediction of general medical admission length of stay with natural language processing and deep learning: a pilot study. Intern Emerg Med 2020 Sep;15(6):989-995. [doi: 10.1007/s11739-019-02265-3] [Medline: 31898204]
4. Chrusciel J, Girardon F, Roquette L, Laplanche D, Duclos A, Sanchez S. The prediction of hospital length of stay using unstructured data. BMC Med Inform Decis Mak 2021 Dec 18;21(1):351. [doi: 10.1186/s12911-021-01722-4] [Medline: 34922532]
5. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med 2018 May 18;1:18. [doi: 10.1038/s41746-018-0029-1] [Medline: 31304302]
6. Bacchi S, Gluck S, Tan Y, et al. Mixed-data deep learning in repeated predictions of general medicine length of stay: a derivation study. Intern Emerg Med 2021 Sep;16(6):1613-1617. [doi: 10.1007/s11739-021-02697-w] [Medline: 33728577]
7. Safavi KC, Khaniyev T, Copenhaver M, et al. Development and validation of a machine learning model to aid discharge processes for inpatient surgical care. JAMA Netw Open 2019 Dec 2;2(12):e1917221. [doi: 10.1001/jamanetworkopen.2019.17221] [Medline: 31825503]
8. Zhang X, Yan C, Malin BA, Patel MB, Chen Y. Predicting next-day discharge via electronic health record access logs. J Am Med Inform Assoc 2021 Nov 25;28(12):2670-2680. [doi: 10.1093/jamia/ocab211] [Medline: 34592753]
9. Stone K, Zwiggelaar R, Jones P, Parthaláin NM. A systematic review of the prediction of hospital length of stay: towards a unified framework. PLOS Digit Health 2022 Apr 14;1(4):e0000017. [doi: 10.1371/journal.pdig.0000017] [Medline: 36812502]
10. Douw G, Schoonhoven L, Holwerda T, et al. Nurses' worry or concern and early recognition of deteriorating patients on general wards in acute care hospitals: a systematic review. Crit Care 2015 May 20;19(1):230. [doi: 10.1186/s13054-015-0950-5] [Medline: 25990249]
11. Kim K, Jeong S, Lee K, et al. Metrics for electronic-nursing-record-based narratives: cross-sectional analysis. Appl Clin Inform 2016 Nov 30;7(4):1107-1119. [doi: 10.4338/ACI-2016-07-RA-0119] [Medline: 27901174]
12. Marafino BJ, Boscardin WJ, Dudley RA. Efficient and sparse feature selection for biomedical text classification via the elastic net: application to ICU risk stratification from nursing notes. J Biomed Inform 2015 Apr;54:114-120. [doi: 10.1016/j.jbi.2015.02.003] [Medline: 25700665]
13. Romero-Brufau S, Gaines K, Nicolas CT, Johnson MG, Hickman J, Huddleston JM. The fifth vital sign? Nurse worry predicts inpatient deterioration within 24 hours. JAMIA Open 2019 Aug 28;2(4):465-470. [doi: 10.1093/jamiaopen/ooz033] [Medline: 32025643]
14. Kim K, Han Y, Jeong S, et al. Prediction of postoperative length of hospital stay based on differences in nursing narratives in elderly patients with epithelial ovarian cancer. Methods Inf Med 2019 Dec;58(6):222-228. [doi: 10.1055/s-0040-1705122] [Medline: 32349156]

15. Hilton CB, Milinovich A, Felix C, et al. Personalized predictions of patient outcomes during and after hospitalization using artificial intelligence. NPJ Digit Med 2020 Apr 3;3:51. [doi: 10.1038/s41746-020-0249-z] [Medline: 32285012]
16. Choi E, Bahadori MT, Sun J, et al. RETAIN: an interpretable predictive model for healthcare using reverse time attention mechanism. Presented at: NIPS'16: Proceedings of the 30th International Conference on Neural Information Processing Systems; Dec 5-10, 2016; Barcelona, Spain p. 3512-3520.
17. Markus AF, Kors JA, Rijnbeek PR. The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. J Biomed Inform 2021 Jan;113:103655. [doi: 10.1016/j.jbi.2020.103655] [Medline: 33309898]
18. Steinberg E, Jung K, Fries JA, Corbin CK, Pfohl SR, Shah NH. Language models are an effective representation learning technique for electronic health record data. J Biomed Inform 2021 Jan;113:103637. [doi: 10.1016/j.jbi.2020.103637] [Medline: 33290879]
19. Kang Y, Jia X, Wang K, et al. A clinically practical and interpretable deep model for ICU mortality prediction with external validation. AMIA Annu Symp Proc 2020;2020:629-637. [Medline: 33936437]
20. Yang G, Ye Q, Xia J. Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: a mini-review, two showcases and beyond. Inf Fusion 2022 Jan;77:29-52. [doi: 10.1016/j.inffus.2021.07.016] [Medline: 34980946]
21. AlSaad R, Malluhi Q, Boughorbel S. PredictPTB: an interpretable preterm birth prediction model using attention-based recurrent neural networks. BioData Min 2022 Feb 14;15(1):6. [doi: 10.1186/s13040-022-00289-8] [Medline: 35164820]
22. Wu J, Dong Y, Gao Z, Gong T, Li C. Dual attention and patient similarity network for drug recommendation. Bioinformatics 2023 Jan 1;39(1):btad003. [doi: 10.1093/bioinformatics/btad003] [Medline: 36617159]
23. Rasmy L, Xiang Y, Xie Z, Tao C, Zhi D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ Digit Med 2021 May 20;4(1):86. [doi: 10.1038/s41746-021-00455-y] [Medline: 34017034]
24. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ 2015 Jan 7;350:g7594. [doi: 10.1136/bmj.g7594] [Medline: 25569120]
25. Flamer HE, Christophidis N, Margetts C, Ugoni A, McLean AJ. Extended hospital stays with increasing age: the impact of an acute geriatric unit. Med J Aust 1996 Jan 1;164(1):10-13. [doi: 10.5694/j.1326-5377.1996.tb94100.x] [Medline: 8559087]
26. Marfil-Garza BA, Belaunzarán-Zamudio PF, Gulias-Herrero A, et al. Risk factors associated with prolonged hospital length-of-stay: 18-year retrospective study of hospitalizations in a tertiary Healthcare center in Mexico. PLoS One 2018;13(11):e0207203. [doi: 10.1371/journal.pone.0207203] [Medline: 30408118]
27. Min YH, Park HA, Chung E, Lee H. Implementation of a next-generation electronic nursing records system based on detailed clinical models and integration of clinical practice guidelines. Healthc Inform Res 2013 Dec;19(4):301-306. [doi: 10.4258/hir.2013.19.4.301] [Medline: 24523995]
28. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988 Sep;44(3):837-845. [Medline: 3203132]
29. Detsky ME, Harhay MO, Bayard DF, et al. Discriminative accuracy of physician and nurse predictions for survival and functional outcomes 6 months after an ICU admission. JAMA 2017 Jun 6;317(21):2187-2195. [doi: 10.1001/jama.2017.4078] [Medline: 28528347]
30. Huang K, Gray TF, Romero-Brufau S, Tulsky JA, Lindvall C. Using nursing notes to improve clinical outcome prediction in intensive care patients: a retrospective cohort study. J Am Med Inform Assoc 2021 Jul 30;28(8):1660-1666. [doi: 10.1093/jamia/ocab051] [Medline: 33880557]
31. May C, Sibley A, Hunt K. The nursing work of hospital-based clinical practice guideline implementation: an explanatory systematic review using normalisation process theory. Int J Nurs Stud 2014 Feb;51(2):289-299. [doi: 10.1016/j.ijnurstu.2013.06.019] [Medline: 23910398]
32. Rohde E, Domm E. Nurses clinical reasoning practices that support safe medication administration: an integrative review of the literature. J Clin Nurs 2018 Feb;27(3-4):e402-e411. [doi: 10.1111/jocn.14077] [Medline: 28926146]
33. Walshe N, Ryng S, Drennan J, et al. Situation awareness and the mitigation of risk associated with patient deterioration: a meta-narrative review of theories and models and their relevance to nursing practice. Int J Nurs Stud 2021 Dec;124:104086. [doi: 10.1016/j.ijnurstu.2021.104086] [Medline: 34601204]
34. Schnock KO, Kang MJ, Rossetti SC, et al. Identifying nursing documentation patterns associated with patient deterioration and recovery from deterioration in critical and acute care settings. Int J Med Inform 2021 Sep;153:104525. [doi: 10.1016/j.ijmedinf.2021.104525] [Medline: 34171662]
35. Woo K, Song J, Adams V, et al. Exploring prevalence of wound infections and related patient characteristics in homecare using natural language processing. Int Wound J 2022 Jan;19(1):211-221. [doi: 10.1111/iwj.13623] [Medline: 34105873]
36. Soenksen LR, Ma Y, Zeng C, et al. Integrated multimodal artificial intelligence framework for healthcare applications. NPJ Digit Med 2022 Sep 20;5(1):149. [doi: 10.1038/s41746-022-00689-4] [Medline: 36127417]

37. Ying X. An overview of overfitting and its solutions. J Phys Conference Ser 2019 Mar 2;1168:022022. [doi: 10.1088/1742-6596/1168/2/022022]

Abbreviations

AUC: area under the receiver operating curve
EHR: electronic health record
***ICD-9*:** *International Classification of Diseases, Ninth Revision*
J-P: Jackson-Pratt
LOS: length of stay
PCA: patient-controlled analgesia
RETAIN: Reverse Time Attention
SNUBH: Seoul National University Bundang Hospital

Edited by A Benis; submitted 28.12.22; peer-reviewed by A Mitra, Y Mao; revised version received 02.08.23; accepted 09.08.23; published 19.12.23.

Please cite as:
Han S, Kim YB, No JH, Suh DH, Kim K, Ahn S
Predicting Postoperative Hospital Stays Using Nursing Narratives and the Reverse Time Attention (RETAIN) Model: Retrospective Cohort Study
JMIR Med Inform 2023;11:e45377
URL: https://medinform.jmir.org/2023/1/e45377
doi:10.2196/45377

Extracting Clinical Information From Japanese Radiology Reports Using a 2-Stage Deep Learning Approach: Algorithm Development and Validation

Kento Sugimoto[1], PhD; Shoya Wada[1,2], MD; Shozo Konishi[1], MD, PhD; Katsuki Okada[1], MD, PhD; Shirou Manabe[1,2,], PhD; Yasushi Matsumura[1,3], MD, PhD; Toshihiro Takeda[1], MD, PhD

1

2

3

Corresponding Author:
Kento Sugimoto, PhD

Abstract

Background: Radiology reports are usually written in a free-text format, which makes it challenging to reuse the reports.

Objective: For secondary use, we developed a 2-stage deep learning system for extracting clinical information and converting it into a structured format.

Methods: Our system mainly consists of 2 deep learning modules: entity extraction and relation extraction. For each module, state-of-the-art deep learning models were applied. We trained and evaluated the models using 1040 in-house Japanese computed tomography (CT) reports annotated by medical experts. We also evaluated the performance of the entire pipeline of our system. In addition, the ratio of annotated entities in the reports was measured to validate the coverage of the clinical information with our information model.

Results: The microaveragedF_1-scores of our best-performing model for entity extraction and relation extraction were 96.1% and 97.4%, respectively. The microaveragedF_1-score of the 2-stage system, which is a measure of the performance of the entire pipeline of our system, was 91.9%. Our system showed encouraging results for the conversion of free-text radiology reports into a structured format. The coverage of clinical information in the reports was 96.2% (6595/6853).

Conclusions: Our 2-stage deep system can extract clinical information from chest and abdomen CT reports accurately and comprehensively.

(JMIR Med Inform 2023;11:e49041) doi:10.2196/49041

KEYWORDS

natural language processing; radiology report; information extraction; deep learning; machine learning; radiology; report; reports; NLP; free text; unstructured; named entity recognition; relation extraction

Introduction

Radiology reports are important for radiologists to communicate with referring physicians. The reports include clinical information about observed structures, diagnostic possibilities, and recommendations for treatment plans. Such information is also valuable for various applications such as case retrieval, cohort building, diagnostic surveillance, and clinical decision support. However, since most radiology reports are written in a free-text format, important clinical information is locked in the reports. This format presents major obstacles in secondary use [1,2]. To address this problem, a system for extracting structured information from the reports would be required.

Natural language processing (NLP) has demonstrated potential for improving the clinical workflow and reusing clinical text for various clinical applications [3-5]. Among the various NLP tasks, information extraction (IE) plays a central role in extracting structured information from unstructured texts. IE mainly consists of two steps: (1) the extraction of specified entities such as person, location, and organization from the text and (2) the extraction of semantic relation between 2 entities (eg,*location_of*and*employee_of*) [6,7].

Earlier IE systems mainly used heuristic methods such as dictionary-based approaches and regular expressions [8-10]. To extract clinical information from radiology reports, the Medical Language Extraction and Encoding system [11] and Radiology Analysis tool [12] have been developed. To detect clinical terms, these systems mainly use predefined dictionaries such as the Unified Medical Language System [13] and their customized dictionaries and apply some grammatical rules to present them in a structured format.

The major issues of these systems include the lack of coverage and scalability [14]. A dictionary-based system often fails to detect clinical terms such as misspelled words, abbreviations, and nonstandard terminologies. Building exhaustive dictionaries to enhance the coverage and maintaining them are highly labor-intensive. It is also challenging to apply complicated grammar rules according to the context of the reports. In addition, IE systems based on dictionaries and grammar rules are highly language dependent and do not scale to other languages. The Medical Language Extraction and Encoding system and Radiology Analysis tool only cover English clinical texts and cannot handle non-English clinical texts. Languages other than English, including Japanese, do not have sufficient clinical resources such as the Unified Medical Language System. This has been a major obstacle in developing clinical NLP systems in countries where English is not the official language [15].

Recently, machine learning approaches have been widely accepted in clinical NLP systems [16,17]. Hassanpour and Langlotz [18] used a conditional random field (CRF) [19] for extracting clinical information from computed tomography (CT) reports. They showed that their machine learning model had a superior ability compared to the dictionary-based systems.

Deep learning approaches have drawn a great deal of attention in more recent studies. Cornegruta et al [20] built a bidirectional long short-term memory (BiLSTM) model [21] to extract clinical terms from chest x-ray reports. Miao et al [22] built a BiLSTM model to handle Chinese radiology reports. Both studies reported that deep learning approaches yielded better results than dictionary-based approaches.

Various state-of-the-art deep learning models have been applied to extract named entities [18,20,22]. Clinical systems such as concept extraction can be achieved though extracting named entities alone, whereas the relation extraction step is needed to obtain structured information about concepts and their attributes [23,24]. Extracting comprehensive information in a structured format is desirable when developing a complex system.

Xie et al [25] developed a 2-stage IE system for processing chest CT reports. They exploited a hybrid approach involving deep learning to extract named entities and a rule-based method to organize the detected entities in a structured format. They reported that their deep learning model achieved better performance, whereas the rule-based structuring approach degraded the overall performance, since the rule-based approach could not capture the contextual relations in the reports. Jain et al [26] developed RadGraph, an end-to-end deep learning system for structuring chest x-ray reports. They reported that their schema had a higher report coverage in their corpus.

In this study, we developed a 2-stage deep learning system for extracting clinical information from CT reports. For secondary use of the radiology reports, we believe that our system has some advantages compared with recent related works [18,20,22,25,26]. First, our 2-stage NLP system can represent clinical information in a structured format, which can be challenging when only using an entity extraction approach. Second, although the rule-based approach struggled to extract relations between entities in the reports [25], leveraging state-of-the-art deep learning models leads to superior performance. Third, previous studies [18,20,26] have combined clinical information about factual observations and radiologist interpretations into single entity, even though they have different semantic roles in the context. According to the context, distinct entity types are defined in our information model, which allows it to capture detailed clinical information in the reports. To structure the report more appropriately, we defined distinct entities for 2 different clinical pieces of information.

The rest of this paper is organized as follows. First, an information model was built, mainly comprising observation entities, clinical finding entities, and their modifier entities. Second, a data set was created using in-house CT reports annotated by medical experts. Third, state-of-the-art deep learning models were trained and evaluated to extract the clinical entities and relations. The entire performance of our 2-stage system was also evaluated. Finally, we evaluated the coverage of the clinical information in the CT reports using our information model.

The development of the information model was already reported in our previous study [27]. However, the previous study only focused on extracting entities and did not cover extracting relations between the entities. This study developed a 2-stage system containing entity extraction and relation extraction modules. Furthermore, although the previous study only used chest CT reports, a data set using abdomen CT reports was created in this study to validate the generalizability of our information model and 2-stage system.

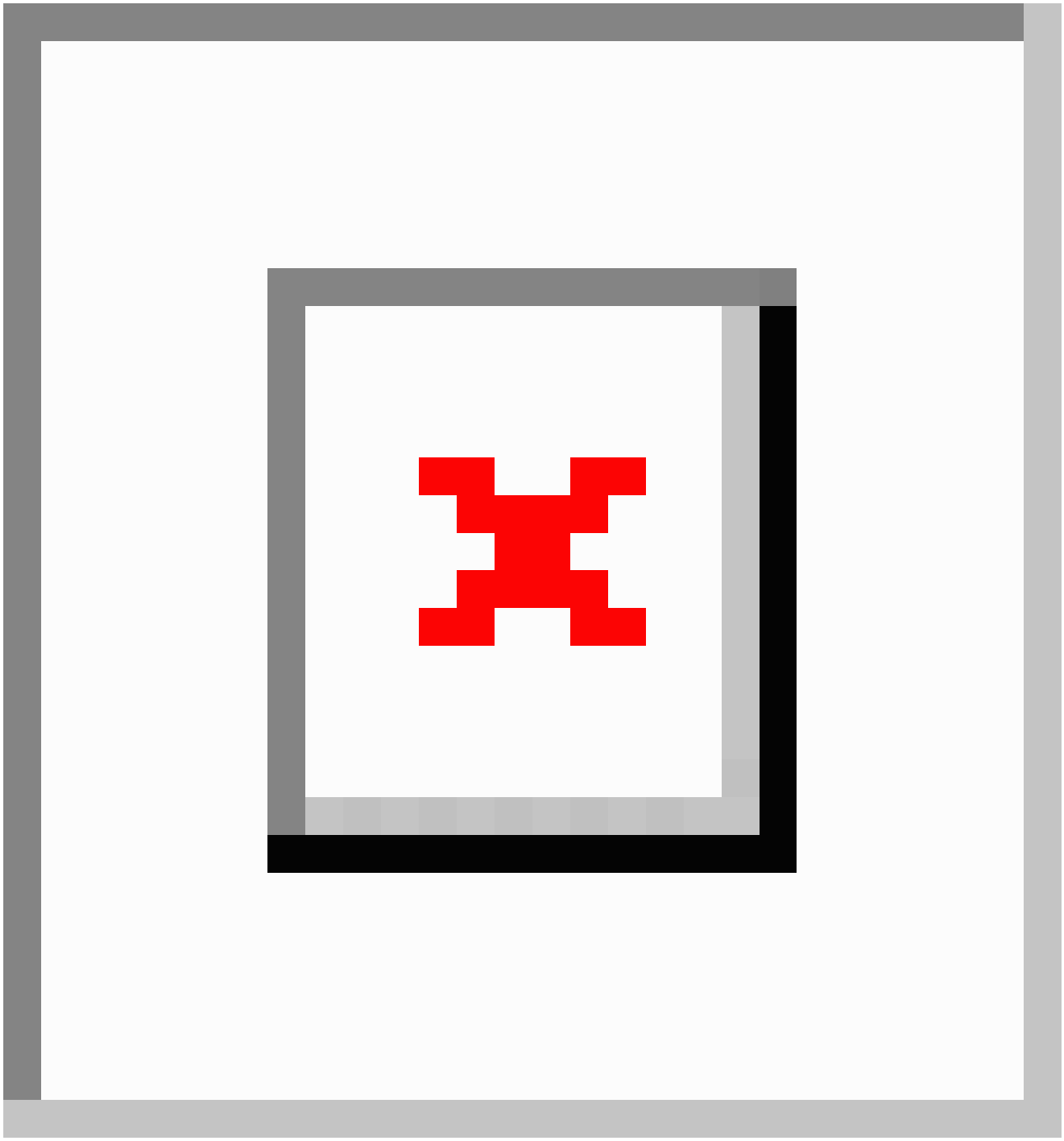
Methods

Our Information Model

An information model was built for extracting comprehensive clinical information from free-text radiology reports. Our information model contained observation entities, clinical finding entities, and modifier entities. Observation entities are specific terms representing observed abnormal features such as “nodule” or “pleural effusion.” Clinical finding entities encompass terms such as “cancer,” including diagnoses given by the radiologists based on the observation entities. Modifier entities are subdivided into the following entities: anatomical location, certainty, change, characteristics, and size. Thus, 7 entity types were defined in our information model. A detailed description of our information model is provided in our previous study [25].

Furthermore, modifier and evidence relations between entities were defined. A modifier relation is derived from an observation or a clinical finding entity and a modifier entity. This relation type gives clinical information, such as the anatomical location of the observations and the characteristics of the clinical findings. An evidence relation is derived from an observation entity and a clinical finding entity. This relation is also clinically meaningful in capturing the diagnostic process of the radiologist. Report examples of entities and relations are shown inFigure 1.

Figure 1. Report examples of entities and relations. IPMN: intraductal papillary mucinous neoplasm.



Data Set

Radiology reports from 2010 to 2021 that were stored in the radiology information system at Osaka University Hospital, Japan, were used. They consisted of 912,505 reports written in Japanese. To create a gold standard data set, 540 chest CT reports and 500 abdomen CT reports were randomly extracted. The remaining unannotated reports (911,465 reports) were used to pretrain the model.

Ethical Considerations

This study was performed in accordance with the World Medical Association Declaration of Helsinki, and the study protocol was approved by the institutional review board of the Osaka University Hospital (permission 19276). Only anonymized data were used in this study, and we did not have access to information that could identify individual participants during the study.

Annotation Scheme

Overall, 3 medical experts (2 clinicians and 1 radiological technologist) performed the annotation process. The gold standard data sets of chest and abdomen CT reports were developed by different annotation methods.

For the chest CT reports, the data set that was developed in our previous study was leveraged [25]. After making minor adjustments for entities, the relation types between entities were newly annotated by 2 clinicians. Following a guideline describing the rules and annotation examples, they

independently annotated each report. Disagreements between the annotators were resolved by discussion. The interannotator agreement (IAA) score for the entities was 91%, as reported in our previous study [27]. To calculate the IAA score for the relations, we used Cohen κ [28], resulting in an IAA score of 81%. Both IAA scores indicated very high agreement [29].

For the abdomen CT reports, to reduce the burden of the annotation work, a deep learning model trained on the chest CT reports was implemented to preannotate the entities and relations in the reports. Annotators were provided with the preannotated reports, and they modified the result according to the guidelines. We did not compute IAA scores for the abdomen data set because it was preannotated by the deep learning model.

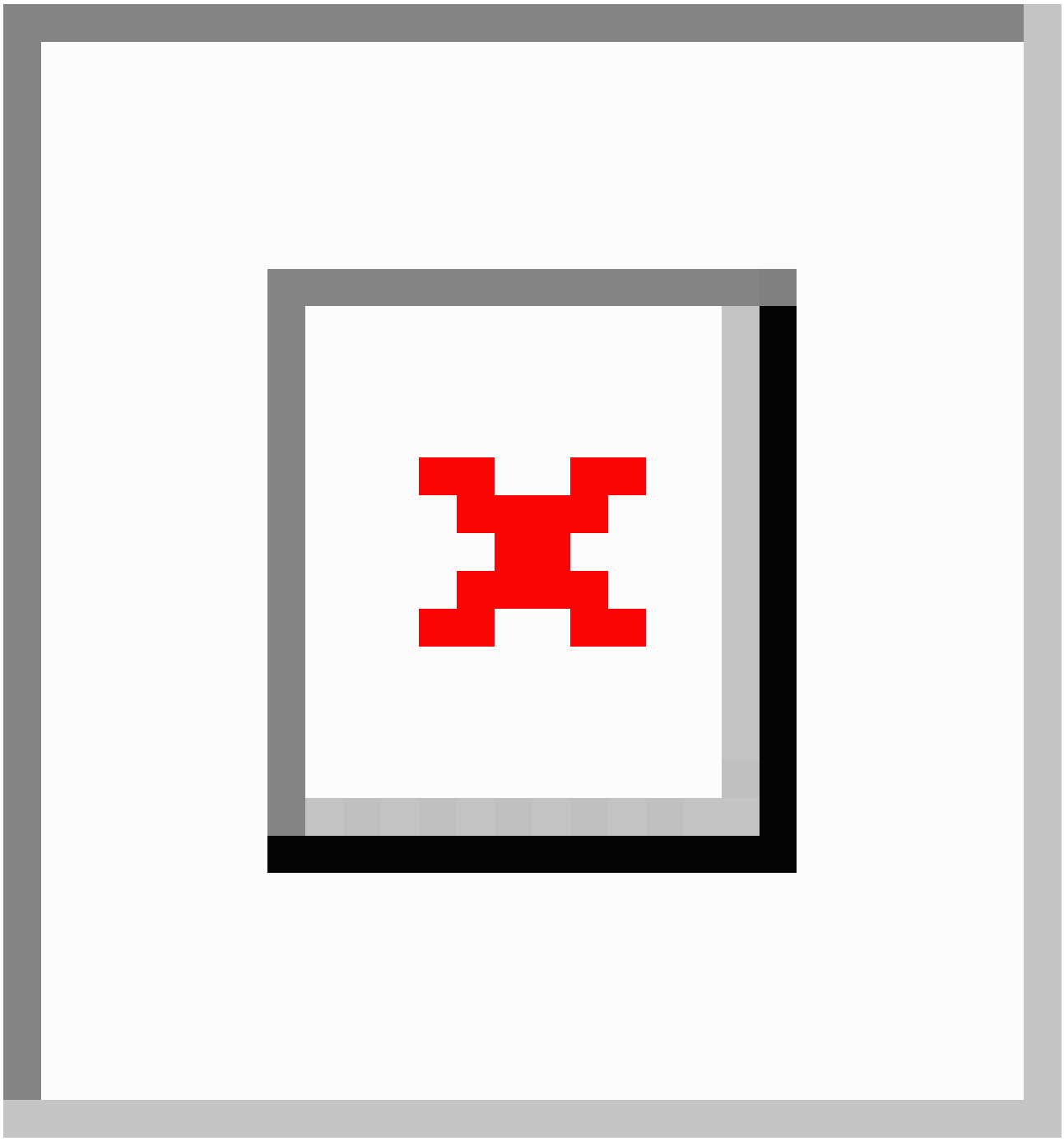
All entities and relations were annotated using BRAT (Stenetorp et al [30]). The number of annotated entities and relations are shown inMultimedia Appendix 1.

Our 2-Stage System

Overview

An overview of our 2-stage system is shown inFigure 2. The system pipeline mainly consists of 2 deep learning modules. In the first step, our module extracts the clinical entities in the radiology reports according to the predefined information model. The extracted entities are fed into subsequent modules. In the second step, the relation between clinical entities is extracted. The details of each module are described in the subsequent sections.

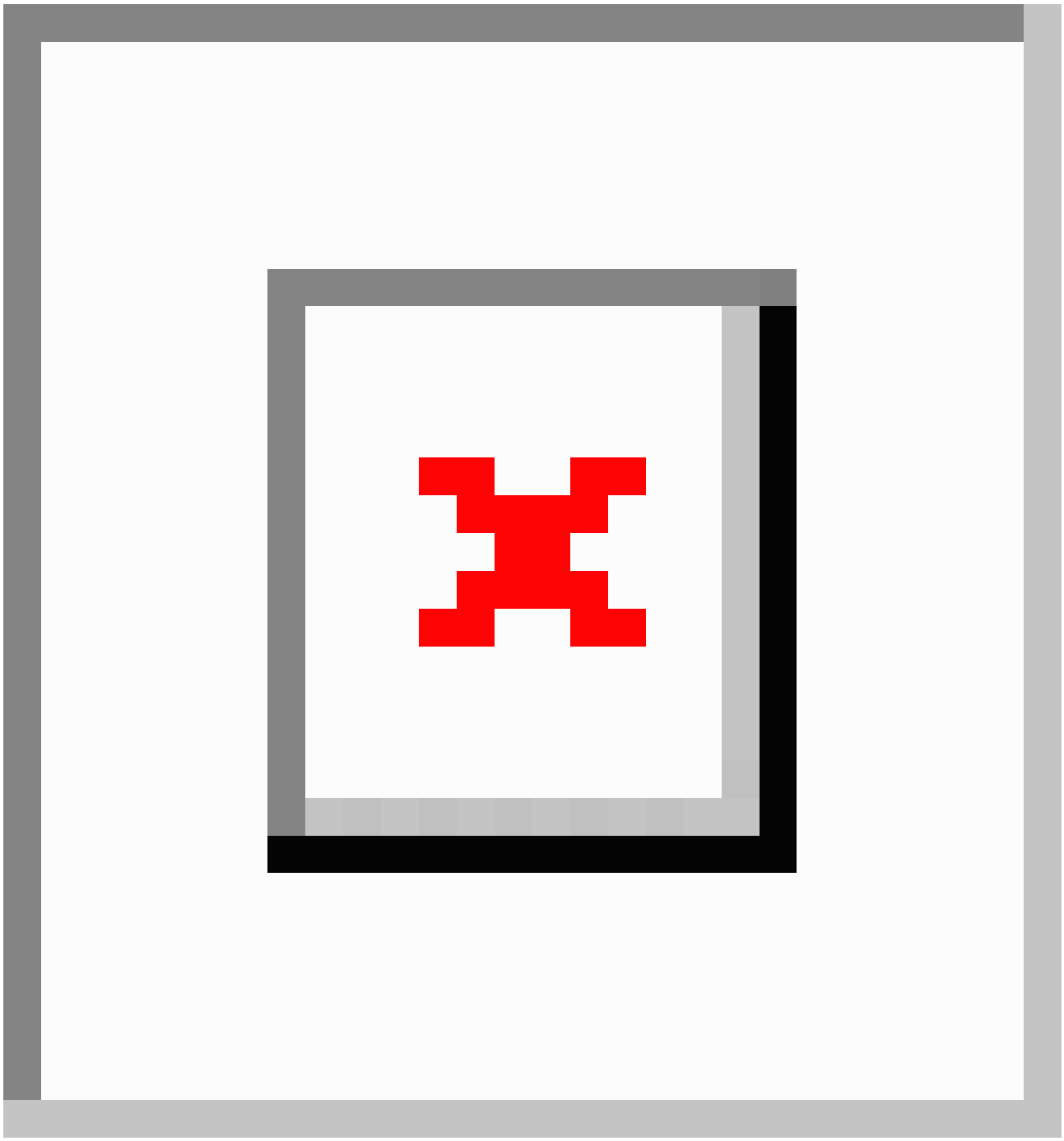
Figure 2. Overview of our 2-stage deep learning system.



Entity Extraction

According to the predefined information model, this module extracts clinical entities from a report. Named entity recognition (NER) [31] is well suited for this task. As a preprocessing pipeline, the report was segmented into sentences using regular expressions, and each sentence was tokenized with MeCab (Kyoto University Graduate School of Informatics and Nippon Telegraph and Telephone Corportation's Communication Science Research Institute) [32]. Then, a sequence of tokens was fed into the model. To represent the spans of specified entities, the IOB2 format [33], which is a widely used tagging format in NER tasks, was used. In this format, the B and I tags represent the beginning and inside of an entity, respectively, and the O tag represents the outside of an entity. A tagging example is illustrated inFigure 3.

Figure 3. An illustration of the entity extraction module. BERT: Bidirectional Encoder Representations from Transformers; BiLSTM: bidirectional long short-term memory; CRF: conditional random field.



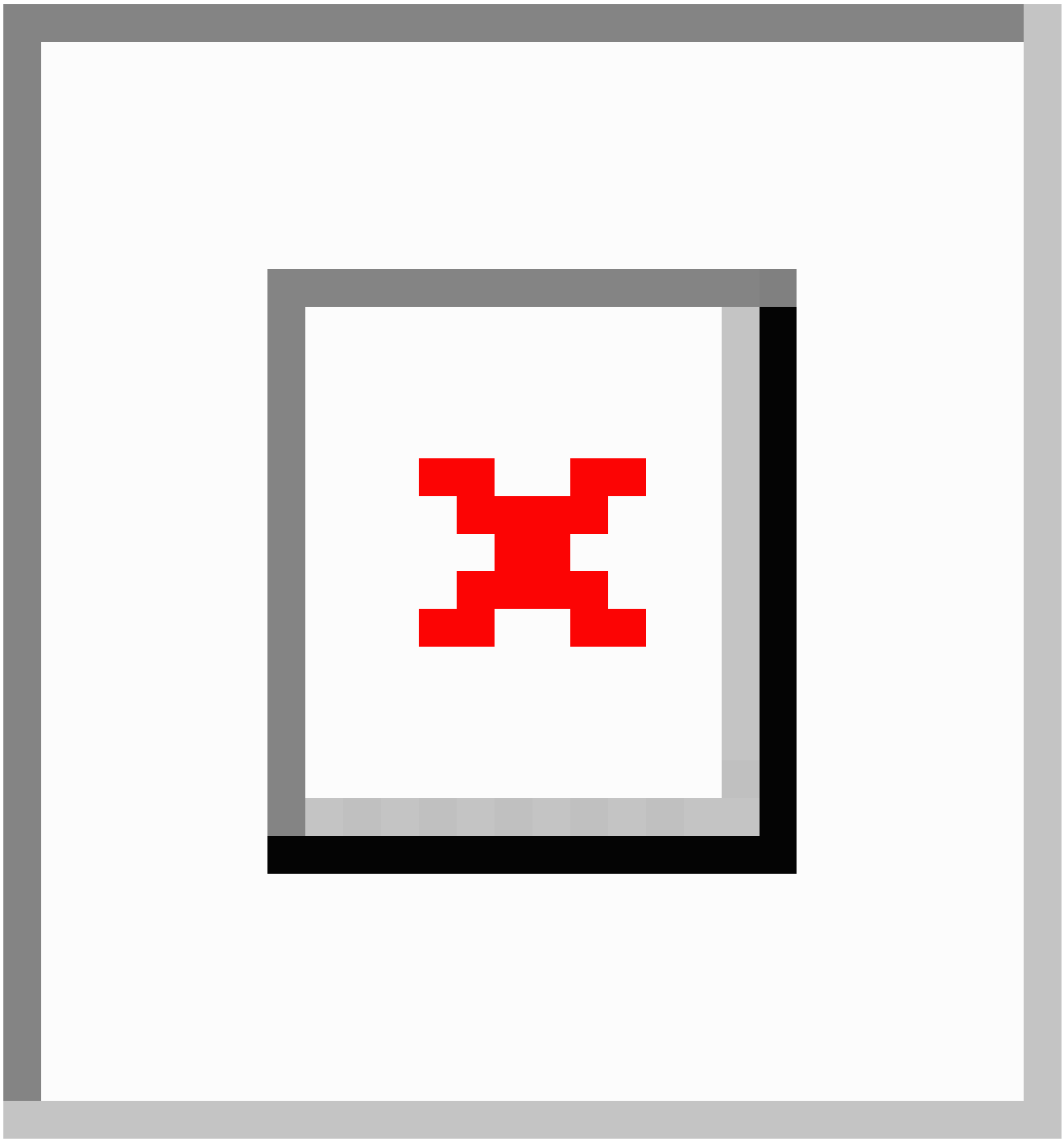
State-of-the-art deep learning models for NER—BiLSTM-CRF [34], BERT [35], and BERT-CRF—were compared.

Relation Extraction

Following the implementation of the entity extraction module, reports with clinical entities were obtained. As a preprocessing pipeline of relation extraction, the original sentences of the report were reconstructed by concatenating sentences from the

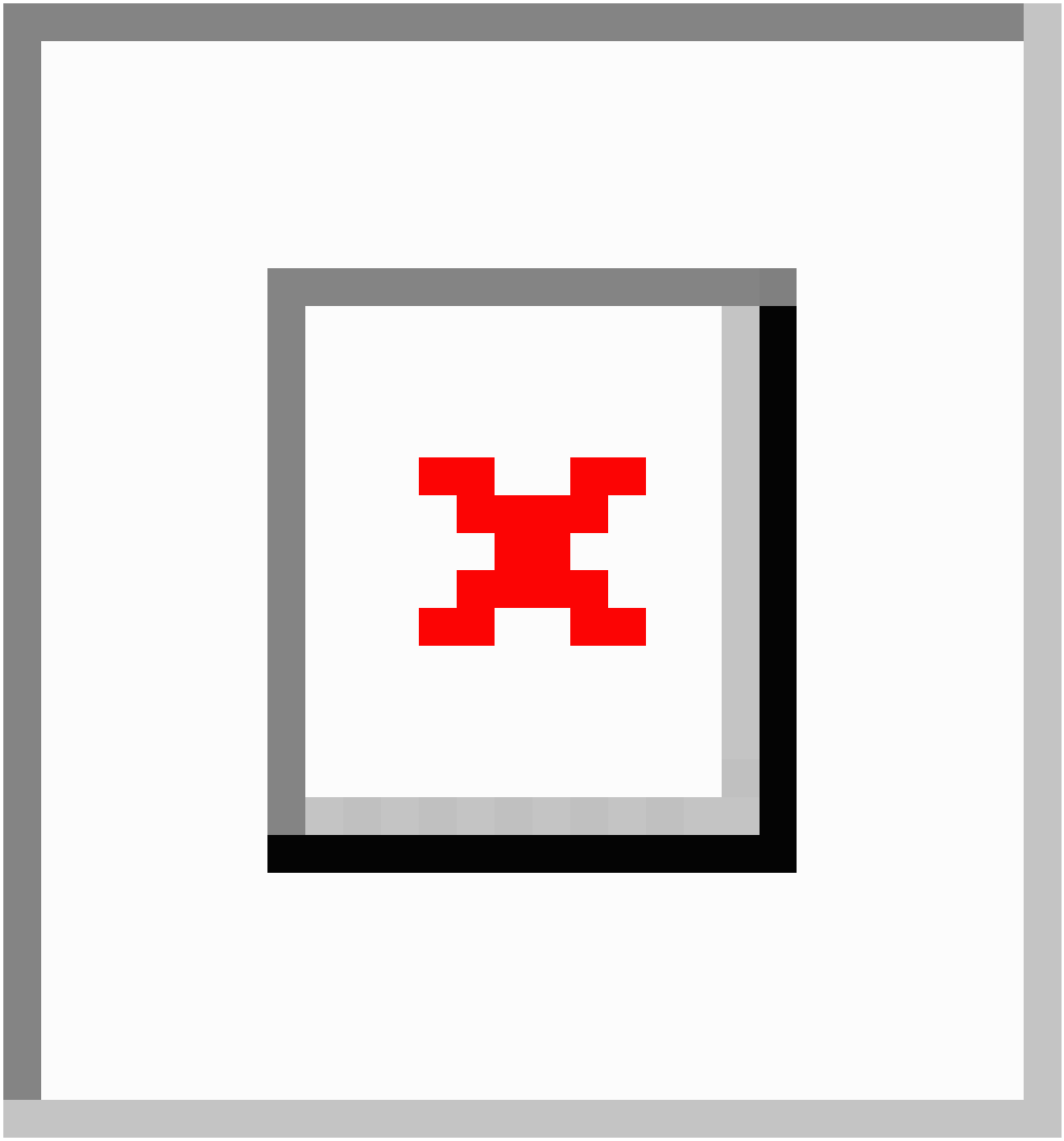
beginning to the end. This was implemented for extracting relations across multiple sentences in a report. Next, the pipeline generated possible candidate relations by each relation type in a report (seeFigure 4). Then, this module solved a binary classification problem to determine the existence of relations given the candidate relations.

Figure 4. Example of instances generated for relation extraction. In this case, 6 candidate relations were generated from 2 observations and 3 modifiers. CT: computed tomography.



Next, we explain how we represented each relation candidate in a fixed-length sequence. Previous studies have introduced a method to add position indicator tokens to the input sequence to indicate the entity span of the pair in the sequence [36,37]. We expanded this method to allow the representation of the entity types. These position indicator tokens are referred to as "entity span tokens." For example, the input sequence of the model representing the relation between an observation entity and an anatomical location modifier entity was represented as follows: "A 3 cm <OBS> nodule </OBS> is in the <AE> right upper lobe </AE>." Here, "<OBS>," "</OBS>," "<AE>," and "</AE>" are entity span tokens. Possible entity span tokens were appended to the vocabulary, and thus, an entity span token was treated as a single token. The input sequence containing 4 entity span tokens was fed into the model. A classification example is illustrated inFigure 5. All generated relation candidates were transformed into fixed-length sequences and fed into the model.

Figure 5. An illustration of the relation extraction module. BERT: Bidirectional Encoder Representations from Transformers; BiLSTM: bidirectional long short-term memory.



The BiLSTM attention model [38] and BERT model were compared. For the BiLSTM attention model, the output vector representation for classification was obtained from the weighted sum of the sequence vector representations. For the BERT model, the representation of the first "[CLS]" token for classification was used, which is a straightforward sequence classification tasks introduced by the original BERT.

Experimental Settings

Data Set Splitting

A total of 540 annotated chest CT reports were divided into 3 groups: 378 reports for training, 54 reports for development, and 108 reports for testing. Similarly, a total of 500 annotated abdomen CT reports were divided into 3 groups: 350 reports for training, 50 reports for development, and 100 reports for testing. In total, 728 reports for training, 94 reports for development, and 208 reports for testing were prepared.

Parameter Optimization

For the BiLSTM-CRF model, a minibatch stochastic gradient descent with momentum was used, and the initial learning rate and momentum were set to 0.1 and 0.9, respectively. The learning rate was reduced when theF_1-score of the development data set stopped improving. Learning rate decay and a gradient clipping of 5.0 were used. Dropout [39] was applied on both the input and output vectors of the BiLSTM model. A batch

size of 16, a dropout rate of 0.1, a word embedding dimension of 100, and a hidden layer dimension of 512 were chosen. For the BERT model, $BERT_{BASE}$was used, which has 12 layers of transformer blocks, 768 hidden units, and 12 self-attention heads. The model was fine-tuned with the initial learning rate of 5×10^{-5}, a batch size of 16, and training epochs of 10. The best hyperparameter setting was chosen using a development data set.

Domain Adaptation

Previous studies have reported that pretraining the domain corpora improved the model performance for various downstream tasks [27,40,41]. However, some studies have pointed out that domain adaptation (DA) leads to a degradation in model performance due to forgetting general domain knowledge [42,43]. To validate the effect of DA in our experiments, we evaluated the model performance with and without DA for both the entity extraction and relation extraction models.

For pretraining the word embeddings of the BiLSTM model with the general domain, Japanese Wikipedia articles [44] (12 million sentences) were used. For pretraining the word embeddings of the BiLSTM model with DA, 911,465 in-house radiology reports were used. We used word2vec (Mikolov et al [45]) for both tasks of pretraining the word embeddings.

For the BERT model, the publicly available pretrained Japanese BERT (Tohoku NLP Group and Tohoku University) [46] was first initialized. The model was pretrained using Japanese Wikipedia articles. The $BERT_{BASE}$subword tokenization model pretrained with whole word masking was chosen. For DA, continued pretraining using 911,465 in-house radiology reports for approximately 100,000 steps using a batch size of 256 was implemented.

Evaluation Metrics

To validate the capability of our system, we conducted 2 experiments. First, the performances of the deep learning modules were calculated. In this experiment, the mean scores were obtained over 5 runs with different parameter initializations to mitigate the effects of a random seed. For both the entity extraction and relation extraction, theF_1-score was used for evaluation. For the entity extraction, entity-levelF_1-score was used as an evaluation metric, and the results were aggregated by microaveraging. Second, to validate that our information model encompassed clinical information in the reports, we measured the coverage with the following formula:

where B-tagged tokens and I-tagged tokens were annotated as entities represented in the IOB2 format [33], and O-tagged tokens as outside entities were not annotated. Following to the scope definition of our information model, the sentences that only contained information about the technique of the imaging test, the surgical procedures of the patients, and recommendations were excluded. Punctuations and stop words were also excluded from the calculation. The list of stop words is presented inMultimedia Appendix 2.

Results

Entity Extraction

Table 1shows the performance metrics for the entity extraction model. The BiLSTM-CRF model with DA achieved a microaveragedF_1-score of 96.1%. In our experiments, the BiLSTM-CRF model with DA achieved the best performance of all the microaveraged scores. For the BERT model, concatenating the CRF layer to the output of the BERT improved the meanF_1-scores with and without DA. Given that the BiLSTM-CRF model with DA yielded the highest meanF_1-score, it was used as the entity extraction module for our system and was used for the remaining experiments.

Table . Comparison of entity extraction models using meanF_1-scores.

Model	Without DA[a], meanF_1-score (%)	With DA, meanF_1-score (%)
BiLSTM[b]	95.2	*96.1*[c]
BERT[d]	94.8	95.2
BERT-CRF[e]	95.1	95.4

[a]DA: domain adaptation.

[b]BiLSTM: bidirectional long short-term memory.

[c]The best performance is italicized.

[d]BERT: Bidirectional Encoder Representations from Transformers.

[e]CRF: conditional random field.

The detailed performance of BiLSTM-CRF model with DA is shown inTable 2. In the test set using chest and abdomen reports, theF_1-scores of observation, clinical finding, anatomical location modifier, certainty modifier, and size modifier entities were over 95%, whereas the change modifier and characteristics modifier entities had lowerF_1-scores than the other entities.Table 2also shows that the test set of abdomen reports had a 0.5% higherF_1-score than the chest reports. On the test set of abdomen reports, the clinical finding and change modifier entities achieved betterF_1-scores than the chest reports, with an increase of 2.9% and 2.5%, respectively. Conversely, the observation and characteristics modifier entities using the test set of chest

reports obtained betterF_1-scores than the abdominal reports, with an increase of 1.0% and 2.6%, respectively.

Table . Comparison of the results of the entity extraction model for the test set of chest and abdomen reports.

Entity type	Chest reports,F_1-score (%)	Abdomen reports,F_1-score (%)	Chest and abdomen reports,F_1-score (%)
Observation	96.1	95.1	95.6
Clinical finding	94.2	97.1	96.1
Anatomical location modifier	96.3	96.3	96.3
Certainty modifier	98.6	99.1	98.9
Change modifier	90.5	93.0	91.5
Characteristics modifier	89.5	86.9	88.5
Size modifier	98.7	98.7	98.7
Microaverage	95.8	96.3	96.1

Relation Extraction

The performances of the relation extraction models were compared. In this experiment, to focus on evaluating the relation extraction module, human-annotated entities were used for the input of each model.Table 3shows the comparisons of the performance of the relation extraction models. A microaveragedF_1-score of 95.6% was achieved for the BiLSTM attention model with DA and 97.6% for the BERT model with DA, which indicated that both classification models could achieve a satisfactory performance for relation extraction. Pretraining with domain corpora improved the performance of both relation models. In contrast to the experimental results of the entity extraction models, the BERT model outperformed the BiLSTM attention model by 2.0% in theF_1-score.

Table . F_1-score of the relation extraction models.

Model	Without DA[a], microaveragedF_1-score (%)	With DA, microaveragingF_1-score (%)
BiLSTM[b]	95.5	*95.6*[c]
BERT[d]	97.2	*97.6*

[a]DA: domain adaptation.

[b]BiLSTM: bidirectional long short-term memory.

[c]The best performance is italicized.

[d]BERT: Bidirectional Encoder Representations from Transformers.

The performance difference between the chest and abdomen reports was also compared (Table 4). TheF_1-scores of the modifier relation were almost the same for the chest reports and abdomen reports, whereas the evidence relation was 6.3% lower in the abdomen reports than the chest reports.

Table . Comparison of the results of the relation extraction model for the test set of chest and abdomen reports.

Relation type and entity type		Chest reports,F_1-score (%)	Abdomen reports,F_1-score (%)	Chest and abdomen reports,F_1-score (%)
Modifier relation				
	Anatomical location	97.9	97.6	97.6
	Certainty	99.4	99.5	99.4
	Change	95.4	95.0	95.1
	Characteristics	95.1	96.5	95.7
	Size	99.1	98.0	98.8
Evidence relation				
	Clinical finding	96.7	90.4	94.9
Microaverage		97.7	97.4	97.6

Our 2-Stage System

To evaluate the performance of the entire pipeline of our system, the performance of the relation extraction module using the output of the entity extraction module was examined. According to the experimental results, the BiLSTM-CRF and BERT models were used for the entity extraction model and relation extraction model, respectively.Table 5shows that the performance of the 2-stage system obtained an overallF_1-score of 91.9%. The overallF_1-score was 5.7% lower than the results using the human-annotated entities, as shown inTable 3. This decrease is reasonable since the misclassification of entity extraction is fed into the relation extraction model in this experiment.

Table . TheF_1-score of our 2-stage system.

Relation type and entity type		2-Stage system,F_1-score (%)
Modifier relation		
	Anatomical location	92.8
	Certainty	96.3
	Change	81.4
	Characteristics	84.7
	Size	94.6
Evidence relation		
	Clinical findings	87.1
Microaverage		91.9

Coverage of Clinical Entities

The test set of reports contained an average of 11.9 sentences. An average of 1.0 (8.4%) out of 11.9 sentences about the technique of the imaging test, the surgical procedures of the patients, and recommendations were excluded from the calculation.Table 6shows the coverage of clinical entities with our information model. The coverage of the clinical entities across entire sequence was 70.2% (7050/10,036). We observed that 96.2% (6595/6853) of tokens were annotated when punctuations and stop words were excluded from the sequences.

Table . Coverage of the clinical entities with our information model.

Token scope	Annotated tokens , n/N (%)
Entire sequence	7050/10,036 (70.2)
Without punctuations and stop words	6595/6853 (96.2)

Error Analysis

A quantitative error analysis was further performed to understand our 2-stage system. For the entity extraction module, we found that the entity mentions that rarely occurred in our corpus were likely missed. To evaluate this empirically, 2 additional test sets were used.

1. Major test set: entity mentions that occurred multiple times in the training set
2. Minor test set: entity mentions that only occured once or did not occur in the training set

Table 7shows the comparison of the result of the major and minor test sets with the original test set (Table 3). In the major test set, theF_1-score of the overall entities was improved by 2.1% (from 96.1% to 98.2%). This increase was also observed in the individual entities except for the size modifier entity. However, theF_1-score of the overall entities was markedly decreased by 9% in the minor test set. This was expected as the deep learning model struggled to predict the samples that were rare or unseen in the training set. Another reason for this difference may be the difficulty in determining the appropriate entities for the minor mentions. We observed that annotation disagreements during the adjudication process occurred more frequently for the minor mentions than the major mentions. Interestingly, we found that the size modifier was robust to the minor entity mentions. The simplicity of these entity mentions, such as “5 cm” and “30×14 mm,” may have contributed to the result. Our analysis shows that the entity extraction module could extract frequent entity mentions in the training set accurately; however, there remains much room for improvement regarding rare or unseen terms in the training set.

Table . Error analysis.

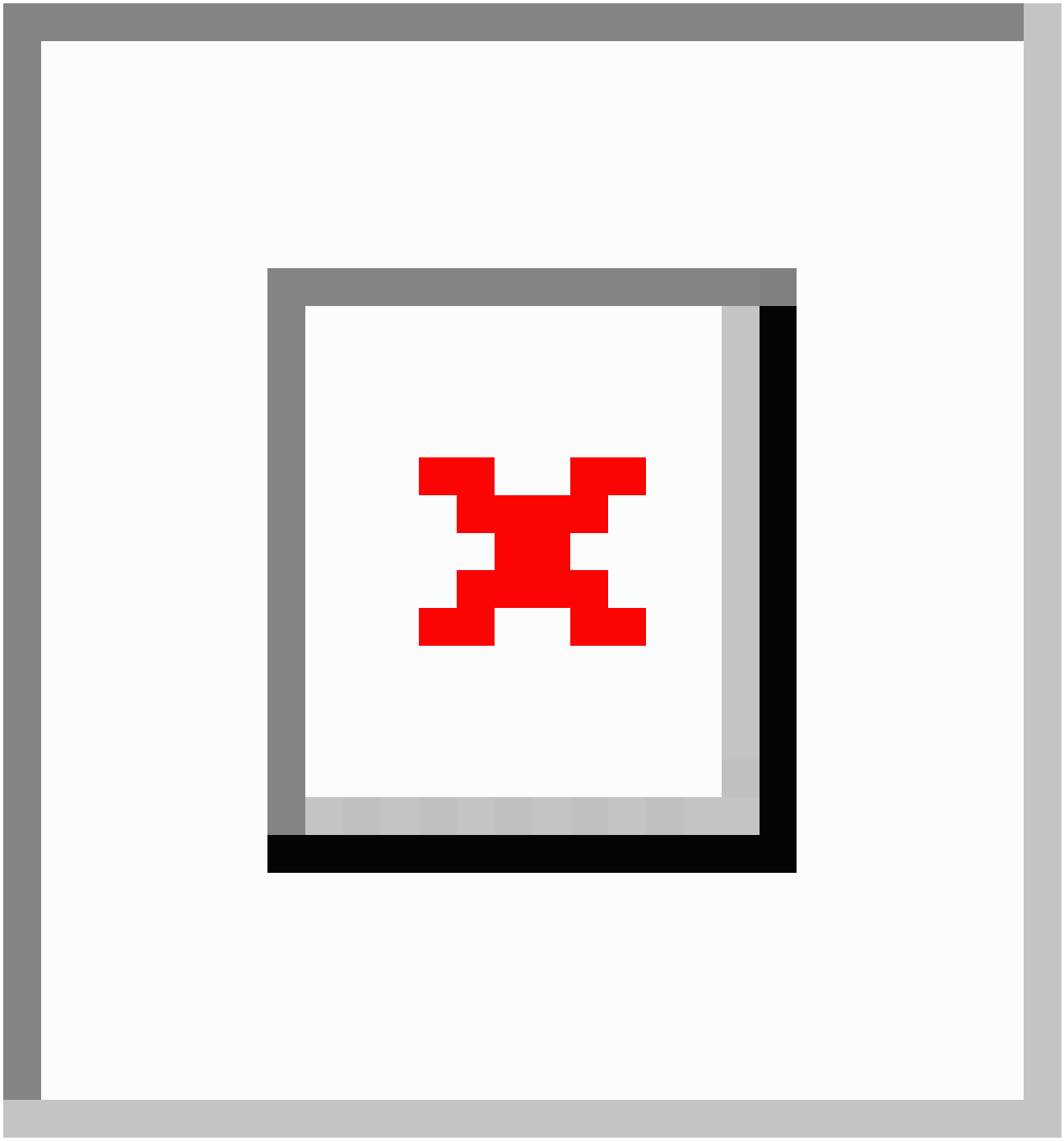
Entity type	Original test set,F_1-score (%)	Major test set		Minor test set	
		F_1-score (%)	Difference from the original test set	F_1-score (%)	Difference from the original test set
Observation	95.6	97.9	+2.3	82.0	–13.6
Clinical finding	96.1	97.9	+1.9	87.8	–8.2
Anatomical location modifier	96.3	98.7	+2.4	89.6	–6.7
Certainty modifier	98.9	99.3	+0.4	80.5	–18.4
Change modifier	91.5	93.5	+2.0	89.0	–2.5
Characteristics modifier	88.5	95.5	+7.1	61.5	–26.9
Size modifier	98.7	98.	–0.3	98.2	–0.6
Microaverage	96.1	98.2	+2.1	87.1	–9.0

To decrease the ratio of rare or unseen terms in the test set, more samples would be required in the training set. However, it is inefficient to sample reports randomly to improve the overall performance. For an efficient sampling strategy, active learning [47,48] may be a promising approach that allows for the selective sampling of reports in the current module.

The performance of the entity extraction and relation extraction modules were compared using the test set of chest and abdomen reports, respectively. For the entity extraction, theF_1-score of the clinical finding entities in the test set for the abdomen reports was 2.9% better than that of the chest reports. In the abdomen reports, it was often written using terms such as “肝臓：n.p. (Liver: n.p.)” when there were no particular findings for a specific organ. This simple expression, “n.p.,” constituted 66.2% of the clinical finding entities in the test set of the abdomen reports, which substantially impacted the performance.

The overall performance of the relation extraction module demonstrated excellent performance on the test set for both the chest and abdomen reports. However, theF_1-score for the evidence relation between the observation and clinical finding entities was 6.3% lower on the test set of the abdomen reports than that of the chest reports. We found a few examples where the observations and clinical findings were clinically related; however, we could not determine if the observation was the diagnostic basis for the finding. The first example shown inFigure 6indicates that the “whirlpool sign” was the observation for the diagnostic basis of an “intestinal obstruction (イレウス),” whereas no observation was found for the diagnostic basis of an “intestinal obstruction (イレウス).” Even though a “whirlpool sign” was clinically related to an “intestinal obstruction (イレウス),” the evidence relation cannot be derived from this example. However, our model misclassified this as a positive example of the evidence relation. In the second example, annotators did not assign the evidence relation between “air” and “biloma,” since they considered that the “air” has already disappeared. However, we discussed that the clinical finding of “biloma” was actually derived from the evidence of an unchanged “low density area (低吸収域)” and disappeared “air.” Thus, the model prediction was more preferable than the gold standard. To derive the diagnostic basis, it is preferable to consider information about the observation and its modifying entities.

Figure 6. Misclassification examples of the relation extraction model (blue highlighted relations are examples of false positives).



Discussion

Principal Findings

Table 3shows the performance of the entity extraction model, which yielded a microaveragedF_1-score of 96.1%. TheF_1-scores of the observation entity and the clinical finding entity were 95.6% and 96.1%, respectively. These superior performances are desirable for our system since the observation and clinical finding entities are principal components of our information model. Moreover,Table 5shows that the modifier relation with the certainty entity also had superior performance. These results suggest that our system will be applicable for practical secondary uses, such as a query-based case retrieval system [49]. However, to reuse radiology reports for various clinical applications, improvements in extracting the change modifier and characteristics modifier would also be required.

BiLSTM Versus BERT

Table 3shows that the BiLSTM-based model achieved better performance than the BERT-based model in the entity extraction task, whereasTable 5shows that the BERT-based model outperformed the BiLSTM-based model in the relation extraction task. We considered that the differences between entity and relation extractions might be due to their task characteristics. Local neighborhood information and the representation of the token itself are considered important in the entity extraction task, whereas more global context

information is required in the relation extraction task, especially for long-distance relations. Due to their attention mechanism, BERT and other transformer-based models are capable of learning long-range dependencies [50], which probably contributed to the superiority of the BERT model in the relation extraction task.

DA Performance

Tables 1and3show the comparison results of the model performances with and without DA for each task. These results indicate that DA is beneficial for performance improvement, regardless of the architecture of the model. Since our system focuses on extracting information from radiology reports, we consider that the problem of forgetting general domain knowledge to be outside the scope of this study.

Coverage of Clinical Entities

The coverage of the clinical entities with our information model was calculated. Sentences about the technique of the imaging test, the surgical procedures of the patients, and recommendations were excluded from the calculation, as such information was outside of the scope of our information model. Punctations and stop words were also excluded from the calculation. A total of 96.2% (6595/6853) of tokens were annotated, which indicates that our information model covered most of the clinical information in the reports.

Limitations

This study has a limitation in terms of generalizability, since we only used 1 institutional data set for evaluation. More data sets outside our institution would be needed to ensure generalizability. Although we validated the capability of our system using only chest and abdomen CT reports, fine-tuning of the deep learning models with reports for other body parts and modalities would be required for various secondary uses.

Furthermore, we are aware that there is still a gap to bridge to reuse radiology reports for various applications. As reports usually contain misspellings, abbreviations, and nonstandard terminologies, we believe that term normalization techniques [51,52] would be needed for clinical applications.

Conclusions

This study developed a 2-stage system to extract structured clinical information from radiology reports. First, we developed an information model and annotated in-house chest and abdomen CT reports. Second, we trained and evaluated the performance of 2 deep learning modules. The microaveragedF_1-scores of our best model for entity extraction and relation extraction were 96.1% and 97.4%, respectively. The entire pipeline of our system achieved a microaveragedF_1-score of 91.9%. Finally, we measured the ratio of annotated entities in the reports. The coverage of the clinical information in the reports was 96.2% (6595/6853). To reuse radiology reports, future studies should focus on term normalization. We also plan to develop a platform that allows us to evaluate the generalizability of our system using reports from outside of our institution.

Acknowledgments

This research was supported by Japan Society for the Promotion of Science KAKENHI grant T22K12885A.

Authors' Contributions

KS developed the entire system, conducted the experiments, and prepared the manuscript. KS, YM, and TT designed the project. YM and TT supervised the project. SW, SK, SM, and KO validated the data. All authors discussed the results and contributed to the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The number of annotated entities and relations.
[PDF File, 94 KB - medinform_v11i1e49041_app1.pdf]

Multimedia Appendix 2
The list of stop words in Japanese.
[PDF File, 148 KB - medinform_v11i1e49041_app2.pdf]

References

1. European Society of Radiology (ESR). ESR paper on structured reporting in radiology. Insights Imaging 2018 Feb;9(1):1-7. [doi: 10.1007/s13244-017-0588-8] [Medline: 29460129]
2. Ganeshan D, Duong PAT, Probyn L, Lenchik L, McArthur TA, Retrouvey M, et al. Structured reporting in radiology. Acad Radiol 2018 Jan;25(1):66-73. [doi: 10.1016/j.acra.2017.08.005] [Medline: 29030284]

3. Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? J Biomed Inform 2009 Oct;42(5):760-772. [doi: 10.1016/j.jbi.2009.08.007] [Medline: 19683066]
4. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet 2012 May 2;13(6):395-405. [doi: 10.1038/nrg3208] [Medline: 22549152]
5. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform 2008;17(1):128-144. [doi: 10.1055/s-0038-1638592] [Medline: 18660887]
6. Sarawagi S. Information extraction. Foundations and Trends in Databases 2008 Nov 30;1(3):261-377. [doi: 10.1561/1900000003]
7. Small SG, Medsker L. Review of information extraction technologies and applications. Neural Comput Appl 2014 Sep;25:533-548. [doi: 10.1007/s00521-013-1516-6]
8. Zeng QT, Goryachev S, Weiss S, Sordo M, Murphy SN, Lazarus R. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. BMC Med Inform Decis Mak 2006 Jul 26;6:30. [doi: 10.1186/1472-6947-6-30] [Medline: 16872495]
9. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc 2010;17(5):507-513. [doi: 10.1136/jamia.2009.001560] [Medline: 20819853]
10. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp 2001:17-21. [Medline: 11825149]
11. Friedman C, Hripcsak G, DuMouchel W, Johnson SB, Clayton PD. Natural language processing in an operational clinical information system. Nat Lang Eng 1995 Mar;1(1):83-108. [doi: 10.1017/S1351324900000061]
12. Johnson DB, Taira RK, Cardenas AF, Aberle DR. Extracting information from free text radiology reports. Int J Digit Libr 1997 Dec;1:297-308. [doi: 10.1007/s007990050024]
13. Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med 1993 Aug;32(4):281-291. [doi: 10.1055/s-0038-1634945] [Medline: 8412823]
14. Taira RK, Soderland SG. A statistical natural language processor for medical reports. Proc AMIA Symp 1999:970-974. [Medline: 10566505]
15. Névéol A, Dalianis H, Velupillai S, Savova G, Zweigenbaum P. Clinical natural language processing in languages other than English: opportunities and challenges. J Biomed Semantics 2018 Mar 30;9(1):12. [doi: 10.1186/s13326-018-0179-8] [Medline: 29602312]
16. Spasic I, Nenadic G. Clinical text data in machine learning: systematic review. JMIR Med Inform 2020 Mar 31;8(3):e17984. [doi: 10.2196/17984] [Medline: 32229465]
17. Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, et al. Clinical information extraction applications: a literature review. J Biomed Inform 2018 Jan;77:34-49. [doi: 10.1016/j.jbi.2017.11.011] [Medline: 29162496]
18. Hassanpour S, Langlotz CP. Information extraction from multi-institutional radiology reports. Artif Intell Med 2016 Jan;66:29-39. [doi: 10.1016/j.artmed.2015.09.007] [Medline: 26481140]
19. Lafferty JD, McCallum A, Pereira FCN. Conditional random fields: probabilistic models for segmenting and labeling sequence data. Presented at: ICML '01: Eighteenth International Conference on Machine Learning; Jun 28-Jul 1, 2001; San Francisco, CA p. 282-289. [doi: 10.5555/645530.655813]
20. Cornegruta S, Bakewell R, Withey S, Montana G. Modelling radiological language with bidirectional long short-term memory networks. Presented at: Seventh International Workshop on Health Text Mining and Information Analysis; Nov 5, 2016; Auxtin, TX p. 17-27. [doi: 10.18653/v1/W16-6103]
21. Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Netw 2005;18(5-6):602-610. [doi: 10.1016/j.neunet.2005.06.042] [Medline: 16112549]
22. Miao S, Xu T, Wu Y, Xie H, Wang J, Jing S, et al. Extraction of BI-RADS findings from breast ultrasound reports in Chinese using deep learning approaches. Int J Med Inform 2018 Nov;119:17-21. [doi: 10.1016/j.ijmedinf.2018.08.009] [Medline: 30342682]
23. Suárez-Paniagua V, Rivera Zavala RM, Segura-Bedmar I, Martínez P. A two-stage deep learning approach for extracting entities and relationships from medical texts. J Biomed Inform 2019 Nov;99:103285. [doi: 10.1016/j.jbi.2019.103285] [Medline: 31546016]
24. Zhang X, Zhang Y, Zhang Q, Ren Y, Qiu T, Ma J, et al. Extracting comprehensive clinical information for breast cancer using deep learning methods. Int J Med Inform 2019 Dec;132:103985. [doi: 10.1016/j.ijmedinf.2019.103985] [Medline: 31627032]
25. Xie Z, Yang Y, Wang M, Li M, Huang H, Zheng D, et al. Introducing information extraction to radiology information systems to improve the efficiency on reading reports. Methods Inf Med 2019 Sep;58(2-03):94-106. [doi: 10.1055/s-0039-1694992] [Medline: 31514210]
26. Jain S, Agrawal A, Saporta A, Truong SQH, Duong DN, Bui T, et al. RadGraph: extracting clinical entities and relations from radiology reports. . Preprint posted online on Aug 29, 2021.. [doi: 10.48550/arXiv.2106.14463]

27. Sugimoto K, Takeda T, Oh JH, Wada S, Konishi S, Yamahata A, et al. Extracting clinical terms from radiology reports with deep learning. J Biomed Inform 2021 Apr;116:103729. [doi: 10.1016/j.jbi.2021.103729] [Medline: 33711545]
28. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960 Apr;20(1):37-46. [doi: 10.1177/001316446002000104]
29. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977 Mar;33(1):159-174. [Medline: 843571]
30. Stenetorp P, Pyysalo S, Topić G, Ohta T, Ananiadou S, Tsujii J. BRAT: a web-based tool for NLP-assisted text annotation. Presented at: Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics; Apr 23-27, 2012; Avignon, France p. 102-107 URL: aclanthology.org/E12-2021 [accessed 2023-10-23]
31. Li J, Sun A, Han J, Li C. A survey on deep learning for named entity recognition. IEEE Trans Knowl Data Eng 2022 Jan;34(1):50-70. [doi: 10.1109/TKDE.2020.2981314]
32. Kudo T. MeCab: yet another part-of-speech and morphological analyzer. GitHub. URL: taku910.github.io/mecab/ [accessed 2021-04-03]
33. Sang EFTK, Veenstra J. Representing text chunks. Presented at: Ninth Conference of the European Chapter of the Association for Computational Linguistics; Jun 8-12, 1999; Bergen, Norway p. 173-179 URL: aclanthology.org/E99-1023 [accessed 2023-10-23]
34. Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C. Neural architectures for named entity recognition. Presented at: 2016 Conference of the North American Chapter of the Association for Computational Linguistics; Jun 12-17, 2016; San Diego, CA p. 260-270. [doi: 10.18653/v1/N16-1030]
35. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. Presented at: 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; Jun 2-7, 2019; Minneapolis, MN p. 4171-4186. [doi: 10.18653/v1/N19-1423]
36. Zhang D, Wang D. Relation classification via recurrent neural network. arXiv. Preprint posted online on Dec 25, 2015.. [doi: 10.48550/arXiv.1508.01006]
37. Zhou P, Shi W, Tian J, Qi Z, Li B, Hao H, et al. Attention-based bidirectional long short-term memory networks for relation classification. Presented at: 54th Annual Meeting of the Association for Computational Linguistics; Aug 7-12, 2016; Berlin, Germany p. 207-212. [doi: 10.18653/v1/P16-2034]
38. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. arXiv. Preprint posted online on May 19, 2014.. [doi: 10.48550/arXiv.1409.0473]
39. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 2014;15(1):1929-1958. [doi: 10.5555/2627435.2670313]
40. Jauregi Unanue I, Zare Borzeshi E, Piccardi M. Recurrent neural networks with specialized word embeddings for health-domain named-entity recognition. J Biomed Inform 2017 Dec;76:102-109. [doi: 10.1016/j.jbi.2017.11.007] [Medline: 29146561]
41. Gururangan S, Marasović A, Swayamdipta S, Lo K, Beltagy I, Downey D, et al. Don't stop pretraining: adapt language models to domains and tasks. Presented at: 58th Annual Meeting of the Association for Computational Linguistics; Jul 5-10, 2020; Online event p. 8342-8360. [doi: 10.18653/v1/2020.acl-main.740]
42. Wiese G, Weissenborn D, Neves M. Neural domain adaptation for biomedical question answering. Presented at: 21st Conference on Computational Natural Language Learning (CoNLL 2017); Aug 3-4, 2017; Vancouver, BC p. 281-289. [doi: 10.18653/v1/K17-1029]
43. Thompson B, Gwinnup J, Khayrallah H, Duh K, Koehn P. Overcoming catastrophic forgetting during domain adaptation of neural machine translation. Presented at: 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; Jun 2-7, 2019; Minneapolis, MN p. 2062-2068. [doi: 10.18653/v1/N19-1209]
44. Index of /jawiki/latest/: jawiki-latest-pages-articles.xml.bz2. Wikipedia. 2023 Jan 3. URL: dumps.wikimedia.org/jawiki/latest/ [accessed 2023-10-27]
45. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv. Preprint posted online on Sep 7, 2013.. [doi: 10.48550/arXiv.1301.3781]
46. Tohoku NLP Group, Tohoku University. Pretrained Japanese BERT models. GitHub. URL: github.com/cl-tohoku/bert-japanese [accessed 2021-03-01]
47. Settles B. Active learning literature survey. University of Wisconsin-Madison. 2009 Jan. URL: minds.wisconsin.edu/handle/1793/60660 [accessed 2023-10-23]
48. Ren P, Xiao Y, Chang X, Huang PY, Li Z, Gupta BB, et al. A survey of deep active learning. ACM Comput Surv 2021 Oct 8;54(9):1-40. [doi: 10.1145/3472291]
49. Pons E, Braun LMM, Hunink MGM, Kors JA. Natural language processing in radiology: a systematic review. Radiology 2016 May;279(2):329-343. [doi: 10.1148/radiol.16142770] [Medline: 27089187]
50. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Presented at: NIPS'17: 31st International Conference on Neural Information Processing Systems; Dec 4-9, 2017; Long Beach, CA p. 6000-6010. [doi: 10.5555/3295222.3295349]

51. Leaman R, Islamaj Dogan R, Lu Z. DNorm: disease name normalization with pairwise learning to rank. Bioinformatics 2013 Nov 15;29(22):2909-2917. [doi: 10.1093/bioinformatics/btt474] [Medline: 23969135]
52. Leaman R, Lu Z. TaggerOne: joint named entity recognition and normalization with semi-Markov models. Bioinformatics 2016 Sep 15;32(18):2839-2846. [doi: 10.1093/bioinformatics/btw343] [Medline: 27283952]

Abbreviations

BERT: Bidirectional Encoder Representations from Transformers
BiLSTM: bidirectional long short-term memory
CRF: conditional random field
CT: computed tomography
DA: domain adaptation
IAA: interannotator agreement
IE: information extraction
NER: named entity recognition
NLP: natural language processing

Edited by J Klann; submitted 16.05.23; peer-reviewed by J Zaghir, M Torii, T Kang; revised version received 25.09.23; accepted 03.10.23; published 14.11.23.

Please cite as:
Sugimoto K, Wada S, Konishi S, Okada K, Manabe S, Matsumura Y, Takeda T
Extracting Clinical Information From Japanese Radiology Reports Using a 2-Stage Deep Learning Approach: Algorithm Development and Validation
JMIR Med Inform 2023;11:e49041
URL: https://medinform.jmir.org/2023/1/e49041
doi:10.2196/49041

Large Language Models for Epidemiological Research via Automated Machine Learning: Case Study Using Data From the British National Child Development Study

Rasmus Wibaek[1], PhD; Gregers Stig Andersen[1], PhD; Christina C Dahm[2], PhD; Daniel R Witte[2,3], PhD; Adam Hulman[2,3], PhD

1

2

3

Corresponding Author:
Adam Hulman, PhD

Abstract

Background: Large language models have had a huge impact on natural language processing (NLP) in recent years. However, their application in epidemiological research is still limited to the analysis of electronic health records and social media data.

Objectives: To demonstrate the potential of NLP beyond these domains, we aimed to develop prediction models based on texts collected from an epidemiological cohort and compare their performance to classical regression methods.

Methods: We used data from the British National Child Development Study, where 10,567 children aged 11 years wrote essays about how they imagined themselves as 25-year-olds. Overall, 15% of the data set was set aside as a test set for performance evaluation. Pretrained language models were fine-tuned using AutoTrain (Hugging Face) to predict current reading comprehension score (range: 0-35) and future BMI and physical activity (active vs inactive) at the age of 33 years. We then compared their predictive performance (accuracy or discrimination) with linear and logistic regression models, including demographic and lifestyle factors of the parents and children from birth to the age of 11 years as predictors.

Results: NLP clearly outperformed linear regression when predicting reading comprehension scores (root mean square error: 3.89, 95% CI 3.74-4.05 for NLP vs 4.14, 95% CI 3.98-4.30 and 5.41, 95% CI 5.23-5.58 for regression models with and without general ability score as a predictor, respectively). Predictive performance for physical activity was similarly poor for the 2 methods (area under the receiver operating characteristic curve: 0.55, 95% CI 0.52-0.60 for both) but was slightly better than random assignment, whereas linear regression clearly outperformed the NLP approach when predicting BMI (root mean square error: 4.38, 95% CI 4.02-4.74 for NLP vs 3.85, 95% CI 3.54-4.16 for regression). The NLP approach did not perform better than simply assigning the mean BMI from the training set as a predictor.

Conclusions: Our study demonstrated the potential of using large language models on text collected from epidemiological studies. The performance of the approach appeared to depend on how directly the topic of the text was related to the outcome. Open-ended questions specifically designed to capture certain health concepts and lived experiences in combination with NLP methods should receive more attention in future epidemiological studies.

(JMIR Med Inform 2023;11:e43638) doi:10.2196/43638

KEYWORDS

natural language processing; deep learning; language model; epidemiology; cohort study; prediction; NLP; prediction model; child development; regression model; large language model; LLM

Introduction

Understanding human language is not a trivial task for machines. Natural language processing (NLP), that is, the analysis of free text with computational methods, has existed as a scientific field for more than half a century [1]. The introduction of large language models was a major leap for the field around the millennium [2]. The essentials of NLP have been reviewed recently with a clinical target audience in mind [3]. Text can be considered as a sequence of characters or words. These linguistic building blocks are referred to as tokens. Once a text is parsed into tokens, a mathematical representation is generated. The most common approach is to use word embeddings—mapping tokens to numerical vectors. These embeddings are trained (assessed in a data-driven manner) on large data sets, and their key feature is that they preserve the relationships between related words. Transformers, introduced in 2017 [4], are currently the most popular underlying model architecture as they excel at

contextualizing words in sentences. The vast amount of easily accessible textual data on the internet represents a massive resource for training language models. As compared to supervised machine learning (ML) approaches where the outcome (target) is available in the training data set, the situation is more complicated with language modeling, where the assignment of labels is not always straightforward. One popular approach, also applied in one of the most influential language models—Bidirectional Encoder Representations from Transformers (BERT) [5], is masked language modeling, that is, masking a certain proportion of words and considering them as outcomes to be predicted based on the preceding sequences of words. Another approach used in the development of BERT is the prediction of the next sentence in a text out of several options. This is a semisupervised training strategy that makes it possible to turn vast amounts of texts, for example, the English-language Wikipedia corpus, into a training set for a language model [5]. Technological advancements in computational tools (eg, graphical processing units and parallelization) have allowed language models to increase massively in size to hundreds of billions of parameters in recent years and have pushed performance closer and closer to human level in various NLP tasks [5-7]. These language models, developed by tech giants or their subsidiaries, are used in search engines, language translators, and auto-correct functions, among others, affecting our everyday lives.

Large language models have a broad scientific potential as well, and with the advent of transfer learning, they are more and more available for those who do not necessarily have the computational resources of tech giants. Transfer learning is the reuse of a pretrained model for a new data set or even a new prediction task that is different from the one it was originally trained for [8]. This approach unlocks the potential of ML for smaller studies by using knowledge representations (in a form of pretrained parameters) learned in large data sets. The significance of the method for NLP was first demonstrated by Howard and Ruder [9], who improved the predictive performance on several NLP benchmarks by ~30% by training a universal language model and reused it for specialized tasks via transfer learning. Even though transfer learning broadens the group of potential users of large language models and deep learning in general, it still requires specialized skills to apply these models. Web services to automate the training and deployment of ML models (automated ML [AutoML]) have been developed to overcome this barrier and unlock the potential in deep learning for researchers without specialized ML skills; however, their use is not common in the clinical research community [10,11].

In addition to knowledge identification (named-entity recognition), synthetization, or discovery in the scientific literature [12], NLP has had an impact on clinical research with applications mostly focusing on the analysis of electronic health records or social media data [13,14], most likely due to the large size of these data sources. However, the potential in free-text data and NLP are to date not fully exploited in classical epidemiological studies. It is likely that NLP performs better than classical regression prediction models in certain settings, but not all, depending on the content of the input text and the outcome to be predicted.

We designed a case study to evaluate the performance of large language models, trained via AutoML, in predicting current reading comprehension and future BMI and physical activity based on essays written by 11-year-old children about how they imagine themselves as 25-year-olds. We then compared this with a classical regression approach, including demographic and lifestyle factors that were selected based on prior domain knowledge as predictors. We explicitly aimed to study and compare the predictive ability of the models (accuracy or discrimination), without the consideration of etiology as it is only on this benchmark that ML and traditional models can currently be compared.

Methods

Data Source

The National Child Development Study (NCDS) originally included 17,415 individuals born in the same week of 1958 in England, Wales, or Scotland [15]. In a total of 12 sweeps, cohort members have been followed since then via interviews, surveys, and biomedical measurements, mostly focusing on health and sociodemographic information not only of the participants but also to some extent their parents. In this study, we used information from baseline (at birth in 1958), sweep 1 (age 7 years in 1965), sweep 2 (age 11 years in 1969), and sweep 5 (age 33 years in 1991) [16-18].

The three outcomes are (1) reading comprehension score (continuous) at age 11 years; (2) BMI (continuous) at age 33 years; and (3) physical activity (binary) at age 33 years. Reading comprehension (score range: 0-35) was assessed using a test filled out at school. The original test is available on the web on the UK Data Service portal [17]. BMI was calculated as weight (kg) divided by height (m) squared based on anthropometric measurements taken at the time of the interview. Physical activity was assessed with 2 questions, asking whether participants do any sport or exercise, and if so, how often. Participants were considered as physically active if they reported exercising at least once a week.

At the age of 11 years, the children were asked to write an essay about how they imagined themselves as 25-year-olds [16]. The instructions were the following: “Imagine that you are 25 years old now. Write about the life you are leading, your interests, your home life and your work at the age of 25. (You have 30 min to do this.).” Out of the 13,669 essays, 10,567 (77.31%) were transcribed [19], which served as the input for the deep learning analyses.

We had access to the following variables that were available at the birth of the participants: sex, ethnicity, birth weight, gestational age at birth, parity, age and BMI of the mother and father, whether the mother spoke English at home, mother’s smoking habit prior to pregnancy, and social class of the head of the household. Moreover, there was information available on the children’s eating habits at age 7 years (appetite and overeating) and BMI, lifestyle (how often they read books, used parks, and did sports activities), and general ability score (similar

to an IQ test) at age 11 years. These variables, selected based on prior knowledge in relation to the outcomes, are only a minor subset of those available in the cohort. Extensive descriptions of the different sweeps of the study are available on the web on the UK Data Service portal [16-18].

Predictive Modeling and Performance Evaluation Strategy

An analytical sample was defined for each of the 3 outcomes. A random sample of approximately 15% of the participants was reserved as a test set in each of the 3 analytical samples before developing the models, and the remaining 85% constituted the development set. In the AutoML approach, the development set was further split into a training set (80%) and a validation set (20%). All reported performance metrics were evaluated on the test sets.

The root mean square error (RMSE) was used as a performance metric for the continuous outcomes, that is, reading comprehension and BMI. Additionally, 95% CIs for RMSE were calculated using the basic bootstrap method with the *boot* package (version 1.3-28) in R (The R Foundation for Statistical Computing). To provide a benchmark RMSE score for comparison, we applied and evaluated a naive approach, that is, assigned the mean value of the outcome from the development data set as predictions in the test set.

Discrimination, measured by the area under the receiver operating characteristics curve (AUC ROC), was used as a performance metric for the binary outcome: physical activity. The naive benchmark was random assignment, and thus, an AUC ROC of 0.5 was defined.

Classical Approach: Regression Models

Regression models included predefined sets of variables that could vary for the 3 outcomes based on prior epidemiological knowledge. Models were fitted using the entire development set after applying multiple imputation (within the development set for each particular outcome) by chained equations to impute missing predictors (*mice* R package; version 3.14.0). We generated 10 imputed samples with the maximum number of iterations set to 30. Estimates were then pooled from the 10 resulting models. The *mice* models derived in the development sets were subsequently applied to the test sets to avoid information leakage.

Reading comprehension score and BMI were modeled using linear regression. For the reading comprehension outcome, we fitted 2 models, with and without including the general ability score among the predictors. The binary outcome physical activity at age 33 years was modeled with logistic regression. The complete list of variables included in each model are shown in Tables 1 and 2.

Table . Linear regression coefficients from prediction models for reading comprehension score and BMI.

Predictor	Reading comprehension score at age 11 years (n=8890)			BMI at age 33 years (n=6010)	
	Imputed, n (%)	Model 1 coefficients (95% CI)	Model 2 coefficients (95% CI)	Imputed, n (%)	Model coefficients (95% CI)
Sex (reference: male)	0	–0.07 (–0.31 to 0.18)	–0.63 (–0.81 to –0.45)	0	–1.1 (–1.3 to –0.9)
Ethnicity (reference: European)	1358 (15.28)			794 (13.21)	
African		–3.20 (–4.41 to –2.00)	–0.60 (–1.41 to 0.22)		1.62 (0.27 to 2.97)
Asian		–2.90 (–4.43 to –1.37)	–1.40 (–2.41 to –0.38)		1.11 (–0.26 to 2.49)
Mother's age (10 years)	423 (4.76)	1.59 (1.23 to 1.95)	0.92 (0.66 to 1.18)	234 (3.89)	–0.22 (–0.56 to 0.12)
Father's age (10 years)	732 (8.23)	0.53 (0.22 to 0.84)	0.37 (0.16 to 0.59)	415 (6.91)	–0.19 (–0.48 to 0.10)
Mother's BMI	N/A[a]	N/A	N/A	615 (10.23)	0.12 (0.09 to 0.14)
Father's BMI	N/A	N/A	N/A	752 (12.51)	0.10 (0.06 to 0.14)
Mother does not speak English at home	1015 (11.42)	–0.06 (–0.46 to 0.34)	–0.14 (–0.42 to 0.14)	N/A	N/A
Birth weight (100 g)	694 (7.81)	0.11 (0.08 to 0.13)	0.03 (0.01 to 0.05)	414 (6.89)	0.00 (–0.03 to 0.02)
Gestational age	1267 (14.25)	0.00 (–0.02 to 0.01)	0.00 (–0.01 to 0.01)	764 (12.71)	0.00 (–0.01 to 0.01)
Parity (reference: nulliparous)	421 (4.74)			232 (3.86)	
Primiparous		–1.33 (–1.64 to –1.03)	–0.72 (–0.94 to –0.49)		–0.12 (–0.38 to 0.14)
Multiparous		–3.45 (–2.81 to –1.48)	–1.53 (–1.78 to –1.27)		0.19 (–0.11 to 0.50)
Maternal smoking	450 (5.06)	–0.41 (–0.66 to –0.17)	0.02 (–0.15 to 0.20)	250 (4.16)	0.38 (0.17 to 0.58)
Socioeconomic status (reference: I)	945 (10.63)			562 (9.35)	
II		–1.24 (–1.85 to –0.63)	–0.27 (–0.72 to 0.18)		–0.30 (–0.80 to 0.20)
III (nonmanual)		–2.14 (–2.81 to –1.48)	–0.56 (–1.05 to –0.08)		0.03 (–0.56 to 0.62)
III (manual)		–4.17 (–4.73 to –3.60)	–1.34 (–1.7 to –0.92)		0.32 (–0.16 to 0.79)
IV		–5.11 (–5.72 to –4.49)	–1.68 (–2.15 to –1.22)		0.26 (–0.27 to 0.79)
V		–6.39 (–7.14 to –5.65)	–1.89 (–2.45 to –1.33)		0.54 (–0.12 to 1.20)
No male head of household		–4.64 (–5.45 to –3.83)	–1.46 (–2.04 to 0.88)		0.27 (–0.39 to 0.93)
Poor appetite	N/A	N/A	N/A	607 (10.1)	–0.15 (–0.43 to 0.13)
Overeating	N/A	N/A	N/A	612 (10.18)	0.06 (–0.39 to 0.52)
Reading books (reference: often)	276 (3.1)			N/A	
Sometimes		–0.61 (–0.86 to –0.36)	–0.20 (–0.39 to –0.02)		N/A
Hardly ever		–2.27 (–2.71 to –1.83)	–0.72 (–1.04 to –0.41)		N/A
Sport (reference: often)	238 (2.68)			146 (2.43)	
Sometimes		0.09 (–0.17 to 0.34)	–0.07 (–0.25 to 0.12)		–0.15 (–0.38 to 0.08)
Hardly ever		–0.06 (–0.47 to 0.35)	0.10 (–0.19 to 0.39)		–0.07 (–0.41 to 0.27)
Park use (reference: often)	847 (9.53)			488 (8.12)	
Sometimes		0.38 (0.12 to 0.65)	0.10 (–0.10 to 0.30)		–0.30 (–0.54 to –0.06)
Never		0.36 (–0.15 to 0.87)	0.32 (–0.04 to 0.69)		0.05 (–0.44 to 0.54)
Not available		0.06 (–0.34 to 0.46)	–0.04 (–0.33 to 0.25)		–0.11 (–0.45 to 0.24)

Predictor	Reading comprehension score at age 11 years (n=8890)			BMI at age 33 years (n=6010)	
	Imputed, n (%)	Model 1 coefficients (95% CI)	Model 2 coefficients (95% CI)	Imputed, n (%)	Model coefficients (95% CI)
General ability score	1 (0.01)	N/A	0.26 (0.26 to 0.27)	N/A	N/A
BMI at age 11 years	N/A	N/A	N/A	886 (14.74)	0.67 (0.62 to 0.72)

[a]N/A: not applicable.

Table . Odds ratios (OR) from the prediction model for physical activity.

Predictor		Outcome: physical activity at age 33 years (n=6204)	
		Imputed, n (%)	OR (95% CI)
Sex (reference: male)		0 (0)	1.13 (1.01-1.27)
Ethnicity (reference: European)		846 (14.04)	
	African		0.68 (0.34-1.35)
	Asian		0.76 (0.40-1.45)
Mother's age (10 years)		234 (3.88)	0.98 (0.83-1.15)
Father's age (10 years)		431 (7.15)	1.05 (0.91-1.21)
Mother's BMI		652 (10.82)	0.98 (0.97-1.00)
Father's BMI		809 (13.43)	0.99 (0.98-1.01)
Birth weight (100 g)		411 (6.82)	1.00 (0.99-1.01)
Gestational age		796 (13.21)	1.00 (0.99-1.01)
Parity (reference: nulliparous)		232 (3.85)	
	Primiparous		0.99 (0.86-1.13)
	Multiparous		0.94 (0.80-1.11)
Maternal smoking		253 (4.2)	0.92 (0.81-1.04)
Socioeconomic status (reference: I)		579 (9.61)	
	II		0.91 (0.69-1.19)
	III (nonmanual)		0.83 (0.61-1.12)
	III (manual)		0.82 (0.63-1.06)
	IV		0.77 (0.59-1.02)
	V		0.64 (0.45-0.92)
	No male head of household		0.72 (0.51-1.02)
Poor appetite		631 (10.47)	0.87 (0.71-1.05)
Overeating		630 (10.46)	0.92 (0.74-1.15)
Sport (reference: often)		148 (2.46)	
	Sometimes		0.90 (0.79-1.02)
	Hardly ever		0.70 (0.59-0.84)
Park use (reference: often)		510 (8.47)	
	Sometimes		0.97 (0.85-1.10)
	Never		0.77 (0.61-0.97)
	Not available		0.73 (0.59-0.90)
BMI at age 11 years		934 (15.50)	1.01 (0.97-1.04)

Deep Learning Approach: NLP Using Large Language Models

We used an AutoML tool, AutoTrain by Hugging Face [20], to develop our NLP prediction model. AutoTrain is a web-based service to train and deploy state-of-the-art ML models (text or tabular as of June 2022). The data sets were uploaded as comma-separated values files including 2 columns: the essays (as text) and the outcome. AutoTrain then split this data set into a training set (80%) and a validation set (20%) and started training (fine-tuning) a variety of pretrained large language models. The number of models can be defined by the user. We chose n=15 for this study. After the training process for all 15 models was complete, we accessed the best-performing model through Hugging Face's application programming interface from Python (Python Software Foundation) and evaluated predictive performance on the reserved 15% in the test set. We did this for all 3 outcomes.

Ethical Considerations

We analyzed a publicly available, anonymized data set; therefore, our study did not require ethical approval.

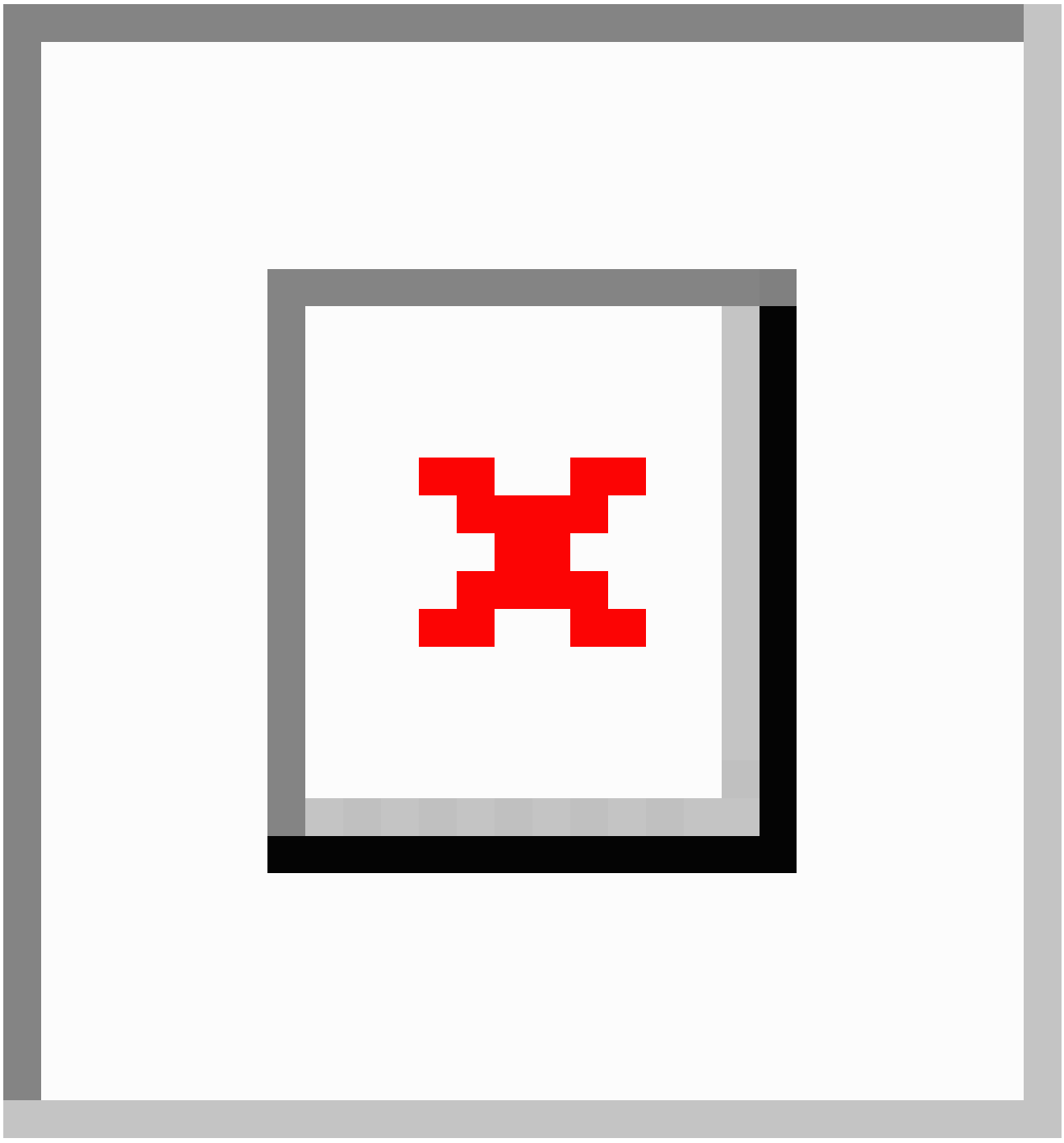
Results

Reading Comprehension Score (Age 11 Years)

Out of 10,567 participants with transcribed essays, 10,490 (99.27%) completed the reading comprehension test, forming the analytical sample for this outcome. From the 10,490 participants, a random sample of 1600 (15.25%) participants were set aside for testing (test set), leaving data from 8890 (84.75%) participants for model development (development set). Reading comprehension scores ranged from 0 to 34, with a median value of 16 (IQR 12-20). The distribution was similar in the test set and only differed slightly in the maximum (35) and the upper-quartile (21) values.

The main results are shown in Figure 1. The naive benchmark had an RMSE of 6.07 (95% CI 5.89-6.26), which was outperformed by both the classical regression and the deep learning approach. The linear regression model without the general ability score had an 11% better performance than the naive benchmark with an RMSE of 5.41 (95% CI 5.23-5.58). This was further improved when including the general ability score in the model (4.14, 95% CI 3.98-4.30). The best performance and thus lowest RMSE was achieved by the deep learning approach (3.89, 95% CI 3.74-4.05), corresponding to a 36% lower RMSE than the naive benchmark.

Figure 1. Performance of the prediction models versus the benchmark approach (naive prediction: assignment of the mean value from the training set) for (A) reading comprehension score and (B) BMI. Root mean square errors (RMSEs) are presented with 95% CIs. Percentages represent differences compared to the benchmark approach. NLP: natural language processing.



The linear regression models revealed that several predictors were associated with the reading comprehension score. Male sex, European ethnicity, having older parents, being the first child in the family, higher birth weight, higher socioeconomic status, reading books often, and having a higher general ability score were all positively associated with reading comprehension. Regression coefficients are presented in Table 1.

BMI (Age 33 Years)

The analytical sample for the BMI analysis consisted of 7060 participants who later had their weight and height measured at age 33 years. From the 7060 participants, a random sample of 1050 (14.87%) participants were set aside for testing model performance, leaving 6010 (85.13%) participants in the development set. BMI values ranged from 12.3 to 50.6 kg/m^2 in the development set and from 15.0 to 50.8 kg/m^2 in the test set. Median values were similar: 24.3 (IQR 22.3-27.1) and 24.4 (IQR 22.2-26.8) kg/m^2, respectively.

Performance metrics are shown in Figure 1. The naive benchmark had an RMSE of 4.45 (95% CI 4.09-4.78), which was similar to the performance of the deep learning approach (4.38, 95% CI 4.02-4.74). The regression model performed ~13% better, achieving an RMSE of 3.85 (95% CI 3.54-4.16).

Several variables were associated with BMI at age 33 years according to the regression model, including sex, ethnicity, parental BMI, parity, maternal smoking before pregnancy, use and access to parks, and BMI at age 11 years. Regression coefficients are presented in Table 1.

Physical Activity (Age 33 Years)

We had information on physical activity at age 33 years from 7304 participants. We selected 1100 (15.06%) of them randomly for the test set, leaving 6204 (84.94%) participants for model development. Overall, 68.75% (4265/6204) and 69.55% (765/1100) were physically active in the development and test sets, respectively. The logistic regression and NLP approaches had the same performance (AUC ROC=0.55, 95% CI 0.52-0.60), representing poor discriminatory power. There were a few variables associated with the outcome in the logistic regression model: sex, socioeconomic status, mother's BMI, sport activities, and use or access to parks at age 11 years. Odds ratios are presented in Table 2.

Discussion

Our study demonstrated the potential of using deep learning–based large language models for text prediction in epidemiological studies and compared it to classical statistical methods. We observed different rankings of predictive performance between the deep learning and classical approaches across the 3 outcomes. The performance of the deep learning approach appeared to depend on how closely the actual task, that is, writing an essay about the future, was related to the outcome. Writing and reading skills among children are expected to be associated with each other, so the language model could have picked up on linguistic features such as grammatical correctness, vocabulary, complexity of sentences, etc, which led to the NLP method clearly outperforming linear regression when predicting the reading comprehension score. This was still true when the general ability score was added to the regression model as a predictor, despite its high correlation with reading comprehension. However, this performance came with a computational price. Large language models include hundreds of millions or even billions of parameters, whereas our regression model included 26. In addition to simplicity, interpretability is another positive feature of linear regression. The model revealed several strong predictors and quantified associations via interpretable regression coefficients, for example, a social gradient with about a 5-point estimated difference between the highest and lowest socioeconomic classes. Although the coefficients are expressed in easily understandable units, they should not be interpreted in the etiological sense, unless a causal framework is applied. With the increasing interest in ML and causal inference, the development of ML methods integrating causal structures is warranted [21].

Epidemiologists and clinicians are comfortable with interpreting the usual measures of association: linear regression coefficients, odds ratios, or hazard ratios. Although we are far from understanding the overall nature of large language models, there are emerging methods in explainable artificial intelligence (AI) that can help to understand the driving factors of at least individual predictions (eg, which features or specific expression in a text led to a prediction). However, they are yet to be integrated into AutoML tools. Access to explainable AI tools (eg, LIME [22]) as part of AutoML solutions is likely to contribute to a more widespread use of deep learning in epidemiological research, where we often ask etiological questions and predictive performance is not necessarily the main focus.

Children were directly asked about their interests as 25-year-olds as part of the essay task, which could potentially include information on physical activity. We therefore expected a similar performance for the NLP and regression approaches. Both approaches picked up some signals in the data, demonstrated by discrimination nominally exceeding random assignment (AUC ROC=0.5), but their performance was still poor and statistically not different from each other. A previous study from the NCDS reported that 42% of boys and 34% of girls mentioned physical activity in their essays [19]. The authors then used this information to predict their physical activity patterns during adulthood, and they found a positive association among boys, but not girls. Pongiglione et al [19] used a 2-step approach: first, they applied a supervised ML method (support vector machines) to extract information on physical activity identity from the essays and, second, used that variable to predict the physical activity in adulthood with a separate logistic regression model. The drawback of this approach is that it needs a subset of the data set to have labels for the intermediate outcome (whether physical activity was mentioned in the text or not), which can be time-consuming and labor-intensive for large data sets. Once some labels are available and the prediction model has reasonable performance, the approach can handle large amounts of data to classify the rest of the essays. We have demonstrated that large language models can be directly applied on the data without first generating new intermediate labels.

The major difference between the study by Pongiglione et al [19] and ours, and in general between many epidemiological and data science approaches, is whether the focus is on the causal understanding of associations (etiology) or on prediction. Although the 2 approaches require different study designs and interpretation, the conflation of etiology and prediction is still common in clinical research (eg, causal interpretation of strong predictors) [23]. Our study showed that despite identifying variables strongly associated with the outcome, overall predictive performance might be poor. Therefore, we should be careful when interpreting and drawing causal conclusions from the results of models developed with a predictive aim and avoid mistakenly stating that altering the level of a component of a predictive model would change the risk of the outcome.

Similar evidence also exists regarding the prevention of obesity. In a meta-analysis of 15 prospective studies, Simmonds et al [24] reported that children or adolescents with obesity were about 5 times more likely to be obese in adulthood than those without obesity. In our study, we also found a strong association between BMI in childhood and adulthood; however, the linear regression model performed only slightly better than the naive benchmark, whereas the NLP approach did not outperform the benchmark at all. We were not surprised that NLP performed worse than regression, considering that these approaches had

matching performance in predicting physical activity, and obesity was not expected to be directly mentioned in the essays, in contrast to physical activity. In general, the results for this outcome strengthen our previous argument that prediction can be difficult even if well-established associations are present at the population level.

The development of prediction models, regardless of the use of ML or classical methods, is not a trivial task (handling of missing data, variable selection, reporting, etc) [25-28]. This is often reflected in the quality of prediction studies and the fact that only a small proportion of published prediction models are actually used in clinical practice [29]. AutoML does not offer a solution for this, as careful study design is still crucial. However, it makes the use of deep learning techniques (including pretrained models) more feasible for epidemiologists, who can use their resources on study design instead of programming tasks. Faes et al [10] recently reported a study where physicians (non-AI experts) achieved similar performance to expert-tuned algorithms in several medical image classification tasks [10]. We only needed to use programming in the NLP analysis to preprocess the essays and for the evaluation of the results, whereas the rest of the process was completed in a browser environment (model evaluation became available in AutoTrain by Hugging Face soon after we finished our analyses). AutoML solutions are often claimed to democratize ML; however, the financial costs are still not negligible. It is indeed a positive development that technical skills and computational resources no longer pose as strong a barrier as before. We should be vigilant that this increased accessibility is accompanied by an increased focus on good study design and research quality. An aspect that AutoML might have a positive influence on is knowledge translation. With the AutoML approach we used, the deep learning model became available right after training and could be used to make predictions for new samples either in the browser or via an application programming interface. The developer can choose to keep the model private or make it public so that the research community can reuse it as a pretrained model, either directly or after fine-tuning, thus potentially leading to multistep, incremental transfer learning.

A major strength of our work is the use of deep learning methods that are currently state of the art in NLP to exploit an innovative data source—in this case, text written by participants in a cohort study. We compared these models with standard methods in epidemiology and discussed similarities and differences between the classical and data science approaches. A strength of deep learning methods in general is the potential reuse of extant trained models. Although the interest in transfer learning is rapidly increasing in clinical research, it is still an almost unknown concept in the epidemiology community, despite some studies demonstrating major benefits, even for tabular data [30,31]. To increase the impact of prediction studies, especially those using ML and deep learning methods, authors should be encouraged to deposit their models on the web and make them openly available. This is a common practice in the data science community, as most developers depend heavily on pretrained deep learning models due to computational requirements. The Hugging Face Model Hub has >50,000 pretrained models, which fits well with the FAIR (findability, accessibility, interoperability, and reusability) principles on reusing digital assets in an open and inclusive manner [32]. In a clinical research setting, even if data accompany publications, which is still rarely the case, sharing resources is almost exclusively restricted to data sets and analysis code.

The children's essays used in our NLP models were not designed to be used for specific prediction tasks. Our main aim was to demonstrate the use of deep learning–based large language models and to compare them to the classical statistical methods used in epidemiological research. In showing that NLP methods can extract features from these texts that are associated with certain traits, our study points toward the potential for extracting meaningful additional data from other extant free-text data sources. Each text data source will have its own historical peculiarities and specific characteristics. In our case, the essays were written half a century ago by children. The practical utility of the presented models outside the context of the UK 1958 birth cohort is consequently likely limited without transfer learning via fine-tuning for adaptation to a new context. It should be noted that language models are usually trained on texts from the internet (eg, Wikipedia) and, as such, mostly represent texts written in the past few decades. Where older texts are included—for example, from older, digitized books—sources will represent texts selected for publication at the time. In all cases, texts written by children are likely to be severely underrepresented in training sets.

A previous review of the clinical literature found no evidence for ML having better predictive performance than traditional statistical methods [33]. Considering the trade-off in the loss of easy interpretability, in most studies, the use of ML does not offer any benefits as long as clinical researchers mostly work with tabular data. However, the integration of new data sources in epidemiological studies (text, medical images, and time series) is only possible by applying deep learning and often transfer learning, which also gives us the opportunity to reuse knowledge between studies. With regard to NLP, large language models have almost achieved human-level performance for various specific tasks; therefore, it may become possible for open-ended questions or essays to replace or at least complement long questionnaires (eg, on diet) in large epidemiological studies. Moreover, NLP offers computational methods, for example, for the analysis of interview transcripts in qualitative studies, which might contribute to closing the gap between qualitative and quantitative research. Byrsell et al [34] analyzed transcribed emergency calls to detect out-of-hospital cardiac arrests using deep learning, and Fagherazzi et al [35] recently gave an overview of the potential of vocal biomarkers (containing both linguistic and acoustic features) in clinical research and practice. With the large-scale collection of such and other novel data types, potentially in combination with tabular data, the role of deep learning in epidemiological research is likely to increase as well. However, we can only exploit its potential and develop high-quality prediction models for clinical or public health use in close collaboration between the data science and clinical research communities.

Acknowledgments

AH is supported by a Data Science Emerging Investigator grant (NNF22OC0076725) by the Novo Nordisk Foundation. DRW and AH are employed at Steno Diabetes Center Aarhus, which is partly funded by a donation from the Novo Nordisk Foundation. The funders had no role in the design of the study. The authors are grateful to all participants of the National Child Development Study and the investigators for making the data openly available.

Data Availability

The data set can be accessed through the UK Data Service after a simple registration process [16-18].

Conflicts of Interest

GSA owns shares in Novo Nordisk A/S. GSA is currently employed by Novo Nordisk A/S. During contribution to the manuscript, GSA was employed by Steno Diabetes Center Copenhagen.

References

1. Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. J Am Med Inform Assoc 2011;18(5):544-551. [doi: 10.1136/amiajnl-2011-000464] [Medline: 21846786]
2. Bengio Y, Ducharme R, Vincent P, Janvin C. A neural probabilistic language model. J Mach Learn Res 2003 Mar 2;3(6):1137-1155. [doi: 10.1162/153244303322533223]
3. Chen PH. Essential elements of natural language processing: what the radiologist should know. Acad Radiol 2020 Jan;27(1):6-12. [doi: 10.1016/j.acra.2019.08.010] [Medline: 31537505]
4. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Presented at: 31st International Conference on Neural Information Processing Systems; June 12, 2017; Long Beach, CA p. 6000-6010. [doi: 10.5555/3295222.3295349]
5. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep Bidirectional transformers for language understanding. Presented at: 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT; June 2-7, 2019; Minneapolis, MN p. 4171-4186. [doi: 10.18653/v1/N19-1423]
6. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. Presented at: 34th International Conference on Neural Information Processing Systems; December 6–12, 2020; Vancouver, BC p. 1877-1901. [doi: 10.5555/3495724.3495883]
7. Rae JW, Borgeaud S, Cai T, Millican K, Hoffmann J, Song F, et al. Scaling language models: methods, analysis & insights from training gopher. arXiv. Preprint posted online on December 1, 2021. . [doi: 10.48550/arXiv.2112.11446]
8. Ebbehoj A, Thunbo M, Andersen OE, Glindtvad MV, Hulman A. Transfer learning for non-image data in clinical research: a scoping review. PLOS Digit Health 2022 Feb 17;1(2):e0000014. [doi: 10.1371/journal.pdig.0000014] [Medline: 36812540]
9. Howard J, Ruder S. Universal language model fine-tuning for text classification. Presented at: 56th Annual Meeting of the Association for Computational Linguistics (Volume 1); July 15–20, 2018; Melbourne, Australia p. 328-339. [doi: 10.18653/v1/P18-1031]
10. Faes L, Wagner SK, Fu DJ, Liu X, Korot E, Ledsam JR, et al. Automated deep learning design for medical image classification by health-care professionals with no coding experience: a feasibility study. Lancet Digit Health 2019 Sep;1(5):e232-e242. [doi: 10.1016/S2589-7500(19)30108-6] [Medline: 33323271]
11. Wan KW, Wong CH, Ip HF, Fan D, Yuen PL, Fong HY, et al. Evaluation of the performance of traditional machine learning algorithms, convolutional neural network and AutoML Vision in ultrasound breast lesions classification: a comparative study. Quant Imaging Med Surg 2021 Apr;11(4):1381-1393. [doi: 10.21037/qims-20-922] [Medline: 33816176]
12. Bose P, Srinivasan S, Sleeman WC, Palta J, Kapoor R, Ghosh P. A survey on recent named entity recognition and relationship extraction techniques on clinical texts. Appl Sci 2021 Sep;11(18):8319. [doi: 10.3390/app11188319]
13. Kreimeyer K, Foster M, Pandey A, Arya N, Halford G, Jones SF, et al. Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. J Biomed Inform 2017 Sep;73:14-29. [doi: 10.1016/j.jbi.2017.07.012] [Medline: 28729030]
14. Spasic I, Nenadic G. Clinical text data in machine learning: systematic review. JMIR Med Inform 2020 Mar 31;8(3):e17984. [doi: 10.2196/17984] [Medline: 32229465]
15. Power C, Elliott J. Cohort profile: 1958 British birth cohort (National Child Development Study). Int J Epidemiol 2006 Feb;35(1):34-41. [doi: 10.1093/ije/dyi183] [Medline: 16155052]
16. University College London IOE, Centre for Longitundinal Studies. National Child Development Study: Age 11, Sweep 2, “Imagine You Are 25” Essays, 1969. London, United Kingdom: UK Data Service; 1696.
17. University College London IOE, Centre for Longitudinal Studies. National Child Development Study: Childhood Data from Birth to Age 16, Sweeps 0-3, 1958-1974. 3rd edition. : National Children’s Bureau NBTF; 1974.

18. University College London IOE, Centre for Longitundinal Studies. National Child Development Study: Age 33, Sweep 5, 1991. 2nd edition. : City University SSRU; 1991.
19. Pongiglione B, Kern ML, Carpentieri JD, Schwartz HA, Gupta N, Goodman A. Do children's expectations about future physical activity predict their physical activity in adulthood? Int J Epidemiol 2020 Oct 1;49(5):1749-1758. [doi: 10.1093/ije/dyaa131] [Medline: 33011758]
20. Hugging Face. AutoTrain. URL: huggingface.co/autotrain [accessed 2022-10-17]
21. Rieckmann A, Dworzynski P, Arras L, Lapuschkin S, Samek W, Arah OA, et al. Causes of outcome learning: a causal inference-inspired machine learning approach to disentangling common combinations of potential causes of a health outcome. Int J Epidemiol 2022 Oct 13;51(5):1622-1636. [doi: 10.1093/ije/dyac078] [Medline: 35526156]
22. Ribeiro MT, Singh S, Guestrin C. ""Why should I trust you?": explaining the predictions of any Classifier". Presented at: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 13, 2016; San Francisco, CA p. 1135-1144. [doi: 10.1145/2939672.2939778]
23. Ramspek CL, Steyerberg EW, Riley RD, Rosendaal FR, Dekkers OM, Dekker FW, et al. Prediction or causality? a scoping review of their conflation within current observational research. Eur J Epidemiol 2021 Sep;36(9):889-898. [doi: 10.1007/s10654-021-00794-w] [Medline: 34392488]
24. Simmonds M, Llewellyn A, Owen CG, Woolacott N. Predicting adult obesity from childhood obesity: a systematic review and meta-analysis. Obes Rev 2016 Feb;17(2):95-107. [doi: 10.1111/obr.12334] [Medline: 26696565]
25. Hemingway H, Croft P, Perel P, Hayden JA, Abrams K, Timmis A, et al. Prognosis research strategy (PROGRESS) 1: a framework for researching clinical outcomes. BMJ 2013 Feb 5;346(1):e5595. [doi: 10.1136/bmj.e5595] [Medline: 23386360]
26. Riley RD, Hayden JA, Steyerberg EW, Moons KGM, Abrams K, Kyzas PA, et al. Prognosis research strategy (PROGRESS) 2: prognostic factor research. PLOS Med 2013 Feb 5;10(2):e1001380. [doi: 10.1371/journal.pmed.1001380] [Medline: 23393429]
27. Steyerberg EW, Moons KGM, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis research strategy (PROGRESS) 3: prognostic model research. PLoS Med 2013 Feb 5;10(2):e1001381. [doi: 10.1371/journal.pmed.1001381] [Medline: 23393430]
28. Hingorani AD, van der Windt DA, Riley RD, Abrams K, Moons KGM, Steyerberg EW, et al. Prognosis research strategy (PROGRESS) 4: stratified medicine research. BMJ 2013 Feb 5;346:e5793. [doi: 10.1136/bmj.e5793] [Medline: 23386361]
29. Bouwmeester W, Zuithoff NPA, Mallett S, Geerlings MI, Vergouwe Y, Steyerberg EW, et al. Reporting and methods in clinical prediction research: a systematic review. PLoS Med 2012;9(5):e1001221. [doi: 10.1371/journal.pmed.1001221] [Medline: 22629234]
30. Desautels T, Calvert J, Hoffman J, Mao Q, Jay M, Fletcher G, et al. Using transfer learning for improved mortality prediction in a data-scarce hospital setting. Biomed Inform Insights 2017 Jun;9:1178222617712994. [doi: 10.1177/1178222617712994] [Medline: 28638239]
31. Gao Y, Cui Y. Deep transfer learning for reducing health care disparities arising from biomedical data inequality. Nat Commun 2020 Oct 12;11(1):5131. [doi: 10.1038/s41467-020-18918-3] [Medline: 33046699]
32. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR guiding principles for scientific data management and stewardship. Sci Data 2016 Mar 15;3:160018. [doi: 10.1038/sdata.2016.18] [Medline: 26978244]
33. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol 2019 Jun;110:12-22. [doi: 10.1016/j.jclinepi.2019.02.004] [Medline: 30763612]
34. Byrsell F, Claesson A, Ringh M, Svensson L, Jonsson M, Nordberg P, et al. Machine learning can support dispatchers to better and faster recognize out-of-hospital cardiac arrest during emergency calls: a retrospective study. Resuscitation 2021 May;162:218-226. [doi: 10.1016/j.resuscitation.2021.02.041] [Medline: 33689794]
35. Fagherazzi G, Fischer A, Ismael M, Despotovic V. Voice for health: the use of vocal biomarkers from research to clinical practice. Digit Biomark 2021 Apr 16;5(1):78-88. [doi: 10.1159/000515346] [Medline: 34056518]

Abbreviations

AI: artificial intelligence
AUC ROC: area under the receiver operating characteristic curve
AutoML: automated machine learning
BERT: Bidirectional Encoder Representations from Transformers
FAIR: findability, accessibility, interoperability, and reusability
ML: machine learning
NCDS: National Child Development Study
NLP: natural language processing
RMSE: root mean square error

Edited by C Lovis; submitted 21.10.22; peer-reviewed by A Bjorkelund, N Jiwani, Y Wang; revised version received 29.06.23; accepted 22.07.23; published 19.09.23.

Please cite as:
Wibaek R, Andersen GS, Dahm CC, Witte DR, Hulman A
Large Language Models for Epidemiological Research via Automated Machine Learning: Case Study Using Data From the British National Child Development Study
JMIR Med Inform 2023;11:e43638
URL: https://medinform.jmir.org/2023/1/e43638
doi:10.2196/43638

Predicting Undesired Treatment Outcomes With Machine Learning in Mental Health Care: Multisite Study

Kasper Van Mens[1,2], MSc; Joran Lokkerbol[3], PhD; Ben Wijnen[4], PhD; Richard Janssen[5,6], PhD; Robert de Lange[7], PhD; Bea Tiemens[1,8,9], PhD

1
2
3
4
5
6
7
8
9

Corresponding Author:
Kasper Van Mens, MSc

Abstract

Background: Predicting which treatment will work for which patient in mental health care remains a challenge.

Objective: The aim of this multisite study was 2-fold: (1) to predict patients' response to treatment in Dutch basic mental health care using commonly available data from routine care and (2) to compare the performance of these machine learning models across three different mental health care organizations in the Netherlands by using clinically interpretable models.

Methods: Using anonymized data sets from three different mental health care organizations in the Netherlands (n=6452), we applied a least absolute shrinkage and selection operator regression 3 times to predict the treatment outcome. The algorithms were internally validated with cross-validation within each site and externally validated on the data from the other sites.

Results: The performance of the algorithms, measured by the area under the curve of the internal validations as well as the corresponding external validations, ranged from 0.77 to 0.80.

Conclusions: Machine learning models provide a robust and generalizable approach in automated risk signaling technology to identify cases at risk of poor treatment outcomes. The results of this study hold substantial implications for clinical practice by demonstrating that the performance of a model derived from one site is similar when applied to another site (ie, good external validation).

(JMIR Med Inform 2023;11:e44322) doi:10.2196/44322

KEYWORDS

treatment outcomes; mental health; machine learning; treatment; model; Netherlands; data; risk; risk signaling; technology; clinical practice; model performance

Introduction

Optimizing Health Care Systems

One of the main challenges in designing an efficient health care system is to prevent offering too many resources to some patients and too little to others. In other words, the challenge is to maximize the opportunity for appropriate care at an individual level [1]. The recent strive for precision or personalized medicine aims to improve health care systems by tailoring treatments to patients more effectively. Patients are grouped in terms of their expected treatment response using diagnostic tests or techniques [2]. However, precision medicine remains a challenge in mental health care because treatments are effective *on average*, but it is difficult to predict exactly whom they will work for [3,4]. Stepped care principles provide a framework to allocate limited health care resources and have been proven to be cost-effective for depression and anxiety [5,6]. In stepped care, treatments start with low intensity unless there is a reason to intensify. Such reasons are identified during treatment when there is a lack of confidence in a positive outcome given the current treatment trajectory. To this extent, routine outcome monitoring (ROM) could be used to observe patterns of early treatment response and identify which patients will probably not benefit from their current treatment [7,8].

Identification of Nonresponders

The system can be improved by earlier and more accurate identification of those nonresponders so that patients do not have to endure periods of care in which they do not improve and could potentially lose interest and drop out. On top of that, scarce health care resources are not wasted by engaging in treatment without the desired effect. However, misclassification comes with a cost. Incorrectly classifying patients as needing more intensified treatment results in the unnecessary use of health care resources on patients who would have benefited from a shorter low-intensity treatment. In many Dutch clinics providing basic mental health care, ROM measurements are part of routine care. This raises the question of whether these ROM data could be used to provide accurate prognostic feedback and support a clinician in maximizing the opportunity for appropriate care on the individual level.

Predicting Outcomes With Machine Learning During Treatment

Techniques from the field of machine learning are aimed at making accurate predictions based on patterns in data. Machine learning can help to identify robust, reproducible, and generalizable predictors of treatment response [3,9-11], and has already been used in health care research, for example, in predicting health care costs and outcomes [12-15]. By discovering associations and understanding patterns and trends within the data, machine learning has the potential to improve care. Machine learning permits a finer detection of which patients are at an elevated risk of experiencing persistent poor and costly health outcomes, and may thus give impetus to a more efficient, personalized, and proactive type of mental health care. Inspired by this knowledge, the study aims to use machine learning on ROM data as a feedback device to signal which patients have an elevated risk of a poor response to treatment [16]. However, the use of complex data, and the associated increasingly complex models, challenges researchers to ensure that these models are clinically interpretable rather than a “black box” [17,18].

Independent Validation

After developing a prediction model, it is recommended to evaluate model performance in other clinical data that was not used to develop the model, as mentioned in the TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis) statement. For example, such a validation would require researchers to have access to a similar data set (ie, in terms of predictor variables and outcomes) stemming from a similar population/clinic and compare model performance on this external independent data set (ie, cross-site design). The lack of independent validation is a major limitation of the extant machine learning literature in health care [19]. In a recent review on machine learning for suicide prediction, the majority of studies reviewed split the data into training and testing sets, whereas none of the studies used a cross-site design in which a model was trained using data from one site and evaluated using data from another [20]. Another recent review looking at applications of machine learning algorithms to predict therapeutic outcomes in depression concluded that most studies did not assess out-of-sample estimates of model fit, which limited their generalizability and likely overestimated predictive accuracy [15]. Therefore, the aim of this study was 2-fold: (1) to predict patients’ response to treatment in Dutch basic mental health care using limited commonly available data from routine care and (2) to compare the performance of these machine learning models across three different mental health care organizations in the Netherlands by using clinically interpretable models. By using commonly available data from routine care, the technical implementation of the model in clinical practice would be straightforward.

Methods

Study Design and Data Collection

Data on mental health treatment and outcomes were collected by a data collection protocol. Mental health care sites from 6 regions in the Netherlands were involved. Patients were treated for mild to severe mental health problems, low risk of suicide, or dangerous behavior. The data set consisted of patient records with a completed treatment from 2014 to 2018. A completed treatment in this setting consists of around 5-12 sessions [21]. The protocol consisted of a predefined set of variables with clear definitions and coding for each variable.

For treatment records to be included in this study, the availability of at least the ROM data as well as certain other variables that could be used for predictions was required. As ROM questionnaires are not mandatory in routine care, ROM data were not available for all patients at all measurements. Records were included when ROM data were available at the start of, during, and at the end of treatment. Of the 6 participating regions, 3 had sufficient treatment records (>1000) with nonmissing values and were included in the study (region 1: n=3020; region 2: n=1484; region 3: n=1948). In each region, patients were treated in multiple settings in both urban and rural areas. A set of 26,912 records had to be excluded from the three sites because there was a missing ROM measurement at either the start or end, such that the outcome could not be determined, or there was no measurement during treatment, such that early treatment response patterns could not be determined. To assess the comparability of the included and excluded treatment records in our analysis, a comparison was made regarding age, sex, diagnosis, and baseline severity between both groups (Table 1).

Table . Comparison of patient characteristics between the included and excluded treatment records.

		Included (n=6452)	Excluded (n=26,912)
Sex, n (%)			
	Female	4077 (63.2)	16,872 (62.7)
	Male	2375 (36.8)	10,040 (37.3)
Age category (years), n (%)			
	<30	1978 (30.7)	8671 (32.2)
	30-40	1541 (23.9)	6701 (24.9)
	40-50	1238 (19.2)	5298 (19.7)
	50-60	1154 (17.9)	4119 (15.3)
	≥60	541 (8.4)	2123 (7.9)
Diagnosis group, n (%)			
	Anxiety	2588 (40.1)	9955 (37)
	Depression	2585 (40.1)	10,831 (40.2)
	Other	1279 (19.8)	6126 (22.8)
Total OQ-45.2[a] score baseline, mean (SD)		80.36 (21.18)	80.60 (23.23)

[a]OQ-45.2: Outcome Questionnaire.

Data Description

This study used treatment records, as opposed to patient records. A treatment record was started whenever a patient began treatment within one of the participating centers. As a result, some patients could have multiple treatment records (355/6452, 5.5% of the records were not unique). ROM assessed the development in symptom severity and functioning using the standardized Dutch version of the Outcome Questionnaire (OQ-45.2) [22]. The OQ-45.2 contains three subscales: Symptom Distress, Interpersonal Relations, and Social Role. The psychometric properties of the Dutch OQ-45.2 are adequate [23].

The idea of this study was to support a stepped care framework by predicting, during treatment, undesired outcomes at the end of treatment. These predictions can trigger a reconsideration of the chosen treatment plan to improve the probability of a desired outcome after finishing the treatment. Desired treatment outcomes are highly personal and dependent on the type of treatment and setting. For this study, we choose to define undesired outcomes as nonimprovement. Based on the principles of reliable change [24], we defined nonimprovement as improving less than a medium effect size on the Symptom Distress subscale of the OQ-45.2 [25]. Our study used data from the so-called *basic mental health care* in the Netherlands. Basic mental health care is cost-effective short-term mental health care with an average Cohen *d* effect size of 0.9 [21]. Despite this high effect size, the aim of this short-term treatment of 5-12 sessions is primarily to increase self-direction and get patients back on track without care as soon as possible. In this study, individual treatment goals were unknown, and therefore, it was decided to define nonimprovement as less than a medium effect size. This is a little more than half of the average improvement in this mental health care setting. Our clinical outcome was derived from the observed change in the Symptom Distress scale on the OQ-45.2. Patients with less than half of an SD improvement in symptom severity at the end of treatment were classified as having an “undesired clinical outcome” (called *nonimprovement* henceforth). With the SD of the Symptom Distress subscale in a Dutch clinical population being 16 [23], nonimprovement was defined as a patient not improving at least 8 points on the Symptom Distress subscale of the OQ-45.2.

An early change was defined as the difference in ROM at baseline and the first ROM during treatment. For both the summed scale scores on the OQ-45.2 as well as the individual items, early change variables were created. Besides the ROM data, a set of clinical and demographic variables were included for prediction such as main diagnosis, age, and living condition. The total set consisted of 163 variables, of which 144 were related to the scores on the OQ-45.2 and 19 to the context of the patient.

Modeling and Validation Strategy

The data set was split across all included locations so that models could be trained on a single location and externally validated on each of the other locations. Nonimprovement was predicted for each location separately based on all available predictors using least absolute shrinkage and selection operator (LASSO) models. LASSO was used both to guarantee interpretability for intended model users and to facilitate explicit comparison between prediction models built in different locations. Moreover, as several measures were derived from the same questionnaire, this could have led to multicollinearity between predictors in the data set. LASSO is a technique that has been argued to be able to deal with multicollinearity and still provide stable and interpretable estimators [26]. All numeric variables were centered and scaled.

Using 10-fold cross-validation with 10 repeats, the optimal hyperparameter was determined by considering 100 possible penalty values (ie, λ) between 0.001 and 1000. For the LASSO

with the optimized penalty, the probability threshold was tuned by optimizing F_1-scores over 36 possible probability values between 0.3 and 0.65. The final LASSO model selected for each site was then applied to each of the other sites for model assessment, reporting sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and area under the curve (AUC) using the optimized probability threshold.

Bootstrapping was used to estimate model performance in the site in which the model was built to have an internally validated measure of model performance to compare with the two externally validated measures of model performance by estimating CIs for all performance scores (ie, sensitivity, specificity, PPV, and NPV). The bootstraps were performed by sampling each data set 1000 times with replacement, resulting in 1000 simulated data sets for each site. The final LASSO model of each of the 3 site-specific models was then applied to the bootstrapped data set, resulting in 1000 confusion matrixes per site. Next, the 2.5 percentile and 97.5 percentile for each performance indicator (ie, sensitivity, specificity, PPV, and NPV) were used to determine the 95% CI for each estimate.

All analyses were performed in R (version 4.0.0; R Foundation for Statistical Computing) [27]. The package *caret* was used to build the models [28]. The package *glmnet* was used to perform the LASSO regression [29]. The package *pROC* was used to analyze the AUCs [30].

Ethical Considerations

Since the database was anonymized with statistical disclosure control techniques [31], there was no need for informed consent or approval by a medical ethics committee (Dutch Civil Law, Article 7:458).

Results

The total data set used in the analyses contained information on 6452 treatment records and included anonymized demographic variables, care-related variables, and information about the severity and types of complaints. The characteristics of the patient populations within each site are shown in Table 2. There are notable differences between baseline symptom severity, the distribution of the main diagnosis, and the percentage of patients with a paid job between sites.

Table . Overview of research population (n=6452).

		Region 1 (n=3020)	Region 2 (n=1484)	Region 3 (n=1948)	*P* value
Care-related variables					
Nonimprovement, n (%)		1028 (34.04)	499 (33.63)	577 (29.62)	.003
Treatment duration (days), mean (SD)		145.19 (64.87)	208.00 (78.35)	205.78 (77.52)	<.001
Treatment sessions (n), mean (SD)		9.73 (2.92)	13.15 (4.03)	11.21 (4.34)	<.001
Type and severity of complaints					
Baseline symptom severity score, mean (SD)		51.42 (13.94)	52.16 (13.45)	48.72 (13.65)	<.001
Baseline social role score, mean (SD)		13.76 (5.06)	14.37 (4.97)	13.79 (5.06)	<.001
Baseline interpersonal relations score, mean (SD)		15.29 (6.08)	17.01 (6.50)	15.28 (6.11)	<.001
Baseline total OQ-45[a] score, mean (SD)		80.47 (21.25)	83.54 (20.76)	77.79 (21.07)	<.001
Diagnosis group, n (%)					<.001
	Anxiety	1300 (43)	562 (37.9)	726 (37.3)	
	Depression	1142 (37.8)	568 (38.3)	875 (44.9)	
	Other	578 (19.1)	354 (23.9)	347 (17.8)	
Demographic variables, n (%)					
Sex					.14
	Female	1878 (62.2)	934 (62.9)	1265 (64.9)	
	Male	1142 (37.8)	550 (37.1)	683 (35.1)	
Age category (years)					<.001
	<30	954 (31.6)	505 (34)	519 (26.6)	
	30-40	694 (23)	369 (24.9)	478 (24.5)	
	40-50	577 (19.1)	249 (16.8)	412 (21.1)	
	50-60	556 (18.4)	241 (16.2)	357 (18.3)	
	≥60	239 (7.9)	120 (8.1)	182 (9.3)	
Origin					<.001
	Native	2838 (94)	1 (0.1)	343 (17.6)	
	Immigrant	68 (2.3)	0 (0)	97 (5)	
	Unknown	114 (3.8)	1483 (99.9)	1508 (77.4)	
Marital status					<.001
	Not married	1612 (53.4)	50 (3.4)	969 (49.7)	
	Married	1052 (34.8)	24 (1.6)	747 (38.3)	
	Divorced/widowed	356 (11.8)	8 (0.5)	224 (11.5)	
	Unknown	0 (0)	1402 (94.5)	8 (0.4)	
Living situation					<.001
	Alone	981 (32.5)	35 (2.4)	571 (29.3)	
	With partner	1638 (54.2)	43 (2.9)	1100 (56.5)	
	Child	248 (8.2)	9 (0.6)	186 (9.5)	
	Other	151 (5)	6 (0.4)	83 (4.3)	
	Unknown	2 (0.1)	1391 (93.8)	8 (0.4)	
Paid job					<.001
	Employed	1071 (35.5)	392 (26.4)	536 (27.5)	

		Region 1 (n=3020)	Region 2 (n=1484)	Region 3 (n=1948)	*P* value
	Not employed	1949 (64.5)	831 (56)	1412 (72.5)	
	Unknown	0 (0)	261 (17.6)	0 (0)	

[a]OQ-45.2: Outcome Questionnaire.

The nonzero LASSO coefficients are shown in Table 3. The most important coefficients, in terms of relative coefficient size, were related to early changes in the Symptom Distress subscale of the OQ-45.2, and the change in the total score of the OQ-45.2. The self-blame measurement at the start of treatment was the only other nonzero coefficient at each of the 3 regions. The coefficient for paid employment stands out in the region 1 model, and age had a notable coefficient in regions 1 and 3. Furthermore, the models contained smaller nonzero coefficients that varied between each site (eg, some OQ-45.2 variables were nonzero in some of the models but not in all of the models). The results of the hyperparameter tuning are shown in Table 4. As shown, the threshold to define a positive class was set between 0.30 (region 4) and 0.34 (region 3), with λ varying from 0.02 (region 5) to 0.16 (region 3).

Table . Nonzero least absolute shrinkage and selection operator coefficients of the three models.

		Region 1	Region 2	Region 3
Intercept		–0.59	–0.76	–1.14
Age		0.05	—[a]	0.04
Number of days between referral and first appointment (waiting queue)		—	0.05	—
Employment (paid job)		–0.39	–0.02	—
Nuisance on job (yes, very much)		—	–0.12	—
Work absence (unknown)		—	—	0.11
OQ-45.2[b] start measurement				
	Self-blame	–0.08	–0.01	–0.07
	Feeling week	—	–0.01	—
	Happiness	—	0.05	—
	Disturbing thoughts	–0.11	—	—
	Stomach	—	—	–0.05
	Relationships	–0.01	—	—
	Sadness	–0.03	—	—
OQ-45.2 middle measurement				
	Suicidal thoughts	—	—	0.03
	Enjoyment	—	—	–0.01
	Relationships	–0.07	—	–0.01
OQ-45.2 early change				
	Stamina	—	—	0.01
	Satisfaction in work or school	–0.01	—	–0.05
	Disturbing thoughts	—	—	0.03
	Stomach	—	—	—
	Hearth	0.01	—	—
	Sleeping	0.03	—	—
	Sadness	0.03	—	—
	Relationships	—	–0.02	—
	Headaches	—	—	0.03
SD OQ-45.2 score (change)		0.97	0.81	1.09
Total OQ-45.2 score (change)		0.07	0.15	—

[a]Not applicable.

[b]OQ-45.2: Outcome Questionnaire.

Table . The parameter settings of the three models.

	Lambda	Probability
Model region 1	0.16	0.34
Model region 2	0.03	0.3
Model region 3	0.02	0.32

The performance of the three models is shown in Table 5. Each model (row) has been evaluated internally and two times externally. Each site (columns) has been used three times: one time for internal validation and two times for the external validation of the other models. The diagonal contains the three internal validations. The CIs of the AUCs overlap, which

indicate that there were no significant differences in the overall performances of the models. The AUCs of the three models in the three internal validations were 0.77 (region 2) and 0.80 (regions 1 and 2). The AUCs of the six external validations ranged from 0.77 to 0.80. An overview of the associated confusion matrixes is attached in Multimedia Appendix 1.

Table . Comparison of internally (diagonal) and externally validated results within each site with 1000 bootstrapped CIs for regions 1, 2, and 3.

Metrics		Region 1 validation	Region 2 validation	Region 3 validation
Region 1 model				
	Sensitivity (95% CI)	0.784 (0.760-0.809)	0.762 (0.725-0.800)	0.780 (0.747-0.813)
	Specificity (95% CI)	0.698 (0.676-0.719)	0.647 (0.617-0.676)	0.673 (0.650-0.697)
	Positive predictive value (95% CI)	0.572 (0.545-0.600)	0.522 (0.486-0.560)	0.501 (0.471-0.534)
	Negative predictive value (95% CI)	0.862 (0.846-0.880)	0.843 (0.818-0.868)	0.879 (0.859-0.898)
	AUC[a] (95% CI)	0.799 (0.783-0.816)	0.771 (0.746-0.794)	0.799 (0.778-0.819)
Region 2 model				
	Sensitivity (95% CI)	0.841 (0.818-0.863)	0.824 (0.789-0.856)	0.868 (0.844-0.896)
	Specificity (95% CI)	0.584 (0.563-0.606)	0.586 (0.554-0.615)	0.548 (0.520-0.574)
	Positive predictive value (95% CI)	0.511 (0.486-0.534)	0.502 (0.466-0.533)	0.447 (0.419-0.477)
	Negative predictive value (95% CI)	0.877 (0.860-0.893)	0.868 (0.841-0.892)	0.908 (0.890-0.927)
	AUC (95% CI)	0.782 (0.765-0.799)	0.774 (0.749-0.798)	0.792 (0.772-0.813)
Region 3 model				
	Sensitivity (95% CI)	0.696 (0.667-0.726)	0.673 (0.633-0.716)	0.742 (0.705-0.779)
	Specificity (95% CI)	0.749 (0.730-0.768)	0.726 (0.699-0.754)	0.732 (0.708-0.754)
	Positive predictive value (95% CI)	0.589 (0.561-0.617)	0.554 (0.517-0.596)	0.538 (0.503-0.573)
	Negative predictive value (95% CI)	0.827 (0.809-0.846)	0.814 (0.789-0.841)	0.871 (0.850-0.890)
	AUC (95% CI)	0.787 (0.771-0.803)	0.768 (0.744-0.792)	0.802 (0.782-0.822)

[a]AUC: area under the curve.

Discussion

Evaluation of Three Models at 3 Sites

The aim of this study was to use machine learning to predict which patients would not substantially benefit from treatment across 3 different mental health care organizations in the Netherlands by using clinically interpretable models. This study used a cross-site design in which the performance of a model developed in one site was compared to the model performance on an external independent data set (ie, 3 × 3 cross-site design, as per the TRIPOD statement). Data from ROM, among other clinical and demographic data, were used for the predictions.

Both the AUC of the internal validations of the three models and the corresponding external validations were in the range of 0.77 to 0.80, indicating fair to good model performance [32]. In addition, the CIs of the AUCs overlapped in each of the 9 evaluations, indicating that the performance estimates were robust and likely to be generalizable to different settings. This could be explained by the fact that LASSO regression is known to be less prone to overfitting compared to other machine learning algorithms, and when evaluated with 1000 times bootstrapping, the internal validations give a good indication of overall performance.

All three models generalized well to the other sites. This is an interesting finding and a promising result for the scalability of the implementation of machine learning models. Decentralized data can be gathered, within the boundaries of the General Data Protection Regulation. A model can be developed within the context of one site and then be exported to other sites, even if those other sites differ in certain characteristics. For example, in this research, the 3 sites differed in geographical location from more rural to urban. The patient populations differed, with some significant differences in the distribution of important variables such as main diagnosis, baseline symptom severity, and percentage of patients with paid employment. The data sources differed in the type of electronic health record system used in clinical practice. Despite these substantial differences, we were able to develop three robust machine learning models with acceptable AUCs that could be applied in all 3 settings.

The sensitivity and specificity of the three models were consistent in each of their external validations. There were differences in these metrics between models, mainly caused by a trade-off between sensitivity and specificity when evaluating model performance with metrics from the confusion matrix. The models of regions 1 and 2 were more shifted toward a higher sensitivity and the model of region 3 toward a higher specificity. However, these differences were a shift in the balance rather than an *absolute difference* between the models, as was indicated by the comparable AUCs.

To give some insight into the practical utility of the model, the results can be translated to a hypothetical clinical scenario. Imagine a health care professional with a caseload of 30 patients working in region 2, with a model created in region 1. About 10 of the 30 patients will not improve according to our data (34%). The model is used by the clinician to support the identification of potential nonimproving patients during treatment. With a sensitivity of 0.76 and a specificity of 0.65 (the results of model 1 applied to region 2), 15 patients will be classified as nonimprovers and 15 will be classified as improvers. Among the improvers, 13 of them will actually improve (ie, NPV=0.84), and among the nonimprovers, 8 of them would actually not improve (ie, PPV=0.52). For half of the patients who are classified as nonimprovers, therefore, the discussion would not be necessary at that time. So the question is whether these models are already good enough to actually use in practice. The idea is that when the model indicates that a patient is on track, there is little reason to change treatment. When the model indicates an elevated risk of nonimprovement, the clinician and patient should discuss the situation and adapt treatment plans if necessary. It is therefore important to see such machine learning models not as black-and-white decision tools but as complementary tools in the identification and stratification of patients in need of more or less care.

Predictive Variables

Although this research was aimed at making predictions, rather than explaining relations, we used LASSO regression to inform clinicians about how the algorithm works. In the health care setting, this is important as health care professionals often want to understand which parameters affect and how they contribute to a prediction [33]. By looking at the coefficients of each LASSO model, it can be concluded that the algorithms rely on the variables' early change in the Symptom Distress subscale and the total scores of the OQ-45.2, as well as having a paid job at the start of the treatment and age. In a paper by McMahon [34], several other studies are mentioned in which early symptom improvement, or lack of it, has been associated with psychiatric treatment outcomes. In a study by Lorenzo-Lucas et al [35], being unemployed, among other factors, predicted a lower likelihood of recovery. There were certain individual OQ-45.2 questionnaire items that were associated with nonzero LASSO coefficients. However, these items differed between the sites, and the size of the coefficients were relatively low. We are, therefore, reluctant to generalize findings on these individual OQ-45.2 items, with small nonzero coefficients, to future prediction research.

The high relative importance of the early change variable (ie, in terms of the absolute values of the coefficients) is likely to contribute to good external model validation, as it is a straightforwardly defined predictor that is less likely to be subject to sampling variation. Furthermore, given the high importance of early change in the model, one could even advocate for an alternative simpler predictive model (ie, a "rule of thumb") using early change only (or combined with weaker predictors, eg, age and employment status).

Strengths and Limitations

The main strength of this study is that we used a 3 × 3 cross-site design to develop and evaluate the algorithms, resulting in three models with an independent validation of their performance. In addition, LASSO regression was used, which is a parametric approach, resulting in a prediction model that is still relatively easy to interpret. Moreover, LASSO is less prone to overfitting, which increased the generalizability of the results. Furthermore, with the use of a data protocol with clear data definition descriptions, we could use readily available data from routine care in the Netherlands, meaning that our approach could easily be adopted in other Dutch basic mental health care organizations using ROM (the R scripts to build and validate the models are available on request). This study has a number of limitations that need to be acknowledged. First, we limited our analysis to treatment records with complete data only. In addition, we could not use every variable described in the data protocol because of missing values on these variables in one of the sites. Moreover, we had to exclude a large set of records because of missing data on the OQ-45.2. However, the excluded group of patients did not substantially differ in sex, age, diagnosis, or baseline symptom severity. Nonetheless, we would like to emphasize that our models cannot be directly applied to other patient populations. Second, our data did not contain information on whether the outcome of the ROM had already been used to alter the treatment strategy. This would underestimate the impact of early change, as patients with only minor or no clinical improvements would have been given a possibly more intensive treatment for them to respond to the treatment. Third, although it is difficult to estimate the required sample size for developing a prognostic model, our data had a relatively small sample size [36]. Fourth, this study chose to define an undesired outcome as improving with less than a medium effect size. However, the definition of an undesired outcome is subjective and will differ between different types of treatment settings. Therefore, our definition cannot directly be generalized to other settings, and each research should make an effort to define a relevant undesired outcome for that domain with experts from clinical practice.

This study was performed within the context of a stepped care framework, in which treatment optimization is required during treatment. Our models heavily rely on predictors derived from early change patterns and can, therefore, not be applied at the start of treatment. Other research could analyze which type of predictors are more suited for a matched care framework and to what extent accurate predictions can be made in treatment response.

Conclusion

Machine learning models provide a robust and generalizable approach in automated risk signaling technology to identify cases at risk of poor treatment outcomes. The results of this study hold substantial implications for clinical practice by demonstrating that the performance of a model derived from one site is similar when applied to another site (ie, good external validation). This is a promising result for the scalability of machine learning models developed in single-center studies. Our findings confirm that routine monitoring provides valuable information that can be used in prognostic models to predict treatment outcomes. Such prognostic models can be used as complementary tools for practitioners in a stepped care framework.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Confusion matrix results.
[DOCX File, 16 KB - medinform_v11i1e44322_app1.docx]

References

1. Janssen R, Busschbach J. Op weg naar gepaste geestelijke gezondheidszorg. Economisch Statistische Berichten 2012;97:81-86.
2. Fernandes BS, Williams LM, Steiner J, Leboyer M, Carvalho AF, Berk M. The new field of 'precision psychiatry'.. BMC Med 2017 Apr 13;15(1):80. [doi: 10.1186/s12916-017-0849-x] [Medline: 28403846]
3. Gillan CM, Whelan R. What big data can do for treatment in psychiatry. Curr Opin Behav Sci 2017 Dec;18:34-42. [doi: 10.1016/j.cobeha.2017.07.003]
4. Rush AJ, Trivedi MH, Wisniewski SR, Nierenberg AA, Stewart JW, Warden D, et al. Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: a STAR*D report. Am J Psychiatry 2006 Nov;163(11):1905-1917. [doi: 10.1176/ajp.2006.163.11.1905] [Medline: 17074942]
5. Von Korff M, Tiemens B. Individualized stepped care of chronic illness. West J Med 2000 Feb;172(2):133-137. [doi: 10.1136/ewjm.172.2.133] [Medline: 10693379]
6. van Orden M, Hoffman T, Haffmans J, Spinhoven P, Hoencamp E. Collaborative mental health care versus care as usual in a primary care setting: a randomized controlled trial. Psychiatr Serv 2009 Jan;60(1):74-79. [doi: 10.1176/ps.2009.60.1.74] [Medline: 19114574]
7. Delgadillo J, de Jong K, Lucock M, Lutz W, Rubel J, Gilbody S, et al. Feedback-informed treatment versus usual psychological treatment for depression and anxiety: a multisite, open-label, cluster randomised controlled trial. Lancet Psychiatry 2018 Jun 21;5(7):564-572. [doi: 10.1016/S2215-0366(18)30162-7] [Medline: 29937396]
8. Lutz W, Hofmann SG, Rubel J, Boswell JF, Shear MK, Gorman JM, et al. Patterns of early change and their relationship to outcome and early treatment termination in patients with panic disorder. J Consult Clin Psychol 2014 Apr;82(2):287-297. [doi: 10.1037/a0035535] [Medline: 24447004]
9. Torous J, Baker JT. Why psychiatry needs data science and data science needs psychiatry. JAMA Psychiatry 2016 Jan;73(1):3-4. [doi: 10.1001/jamapsychiatry.2015.2622] [Medline: 26676879]
10. McIntosh AM, Stewart R, John A, Smith DJ, Davis K, Sudlow C, et al. Data science for mental health: a UK perspective on a global challenge. Lancet Psychiatry 2016 Oct;3(10):993-998. [doi: 10.1016/S2215-0366(16)30089-X] [Medline: 27692269]
11. Bzdok D, Meyer-Lindenberg A. Machine learning for precision psychiatry: opportunities and challenges. Biol Psychiatry Cogn Neurosci Neuroimaging 2018 Mar;3(3):223-230. [doi: 10.1016/j.bpsc.2017.11.007] [Medline: 29486863]
12. Chekroud AM, Zotti RJ, Shehzad Z, Gueorguieva R, Johnson MK, Trivedi MH, et al. Cross-trial prediction of treatment outcome in depression: a machine learning approach. Lancet Psychiatry 2016 Mar;3(3):243-250. [doi: 10.1016/S2215-0366(15)00471-X] [Medline: 26803397]
13. Koutsouleris N, Kahn RS, Chekroud AM, Leucht S, Falkai P, Wobrock T, et al. Multisite prediction of 4-week and 52-week treatment outcomes in patients with first-episode psychosis: a machine learning approach. Lancet Psychiatry 2016 Oct;3(10):935-946. [doi: 10.1016/S2215-0366(16)30171-7] [Medline: 27569526]
14. Iniesta R, Malki K, Maier W, Rietschel M, Mors O, Hauser J, et al. Combining clinical variables to optimize prediction of antidepressant treatment outcomes. J Psychiatr Res 2016 Jul;78:94-102. [doi: 10.1016/j.jpsychires.2016.03.016] [Medline: 27089522]
15. Lee Y, Ragguett RM, Mansur RB, Boutilier JJ, Rosenblat JD, Trevizol A, et al. Applications of machine learning algorithms to predict therapeutic outcomes in depression: a meta-analysis and systematic review. J Affect Disord 2018 Dec 1;241:519-532. [doi: 10.1016/j.jad.2018.08.073] [Medline: 30153635]

16. Delgadillo J, de Jong K, Lucock M, Lutz W, Rubel J, Gilbody S, et al. Feedback-informed treatment versus usual psychological treatment for depression and anxiety: a multisite, open-label, cluster randomised controlled trial. Lancet Psychiatry 2018 Jun 21;5(7):564-572. [doi: 10.1016/S2215-0366(18)30162-7] [Medline: 29937396]
17. Graham S, Depp C, Lee EE, Nebeker C, Tu X, Kim HC, et al. Artificial intelligence for mental health and mental illnesses: an overview. Curr Psychiatry Rep 2019 Nov 7;21(11):116. [doi: 10.1007/s11920-019-1094-0] [Medline: 31701320]
18. Freitas AA. Comprehensible classification models. ACM SIGKDD Explorations Newsletter 2014 Mar 17;15(1):1-10. [doi: 10.1145/2594473.2594475]
19. Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal-external, and external validation. J Clin Epidemiol 2016 Jan;69:245-247. [doi: 10.1016/j.jclinepi.2015.04.005] [Medline: 25981519]
20. Kirtley OJ, van Mens K, Hoogendoorn M, Kapur N, de Beurs D. Translating promise into practice: a review of machine learning in suicide research and prevention. Lancet Psychiatry 2022 Mar;9(3):243-252. [doi: 10.1016/S2215-0366(21)00254-6] [Medline: 35183281]
21. van Mens K, Lokkerbol J, Janssen R, van Orden ML, Kloos M, Tiemens B. A cost-effectiveness analysis to evaluate a system change in mental healthcare in the Netherlands for patients with depression or anxiety. Adm Policy Ment Health 2018 Jul;45(4):530-537. [doi: 10.1007/s10488-017-0842-x] [Medline: 29247271]
22. Lambert M, Morton J, Hatfield D, Harmon C, Hamilton S, Shimokawa K. Administration and Scoring Manual for the OQ-45.2 (Outcome Questionniare). 3rd Edition. Orem, UT: American Professional Credentialing Services; 2004.
23. de Jong K, Nugter MA, Polak MG, Wagenborg JEA, Spinhoven P, Heiser WJ. The Outcome Questionnaire (OQ-45) in a Dutch population: a cross-cultural validation. Clin Psychol Psychother 2007 Aug 6;14(4):288-301. [doi: 10.1002/cpp.529]
24. Jacobson NS, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. J Consult Clin Psychol 1991 Feb;59(1):12-19. [doi: 10.1037//0022-006x.59.1.12] [Medline: 2002127]
25. Cohen J. Statistical Power Analysis for the Behavioral Sciences. 2nd Edition. Hillsdale, NJ: Lawrence Earlbaum Associates; 1988. ISBN: 1483276481.
26. Tibshirani R. Regression shrinkage and selection via the lasso. J Royal Stat Soc Ser B Stat Methodology 1996 Jan;58(1):267-288. [doi: 10.1111/j.2517-6161.1996.tb02080.x]
27. R Development Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2008. ISBN: 3-900051-07-0.
28. Kuhn M, Johnson K. Applied Predictive Modeling. New York, NY: Springer; 2013. ISBN: 978-1-4614-6848-6. [doi: 10.1007/978-1-4614-6849-3]
29. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw 2010;33(1):1-22. [Medline: 20808728]
30. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 2011 Mar 17;12(1):77. [doi: 10.1186/1471-2105-12-77] [Medline: 21414208]
31. Meindl MB, Kowarik DIA, Templ PM, Templ M, Meindl B, Kowarik A. Introduction to statistical disclosure control (SDC). International Household Survey Network. 2018. URL: www.ihsn.org/sites/default/files/resources/ihsn-working-paper-007-Oct27.pdf [accessed 2023-07-31]
32. Li F, He H. Assessing the accuracy of diagnostic tests. Shanghai Arch Psychiatry 2018 Jun 25;30(3):207-212. [doi: 10.11919/j.issn.1002-0829.218052] [Medline: 30858674]
33. Hilhorst L, Stappen JVD, Lokkerbol J, Hiligsmann M, Risseeuw AH, Tiemens BG. Patients' and psychologists' preferences for feedback reports on expected mental health treatment outcomes: A discrete-choice experiment. Adm Policy Ment Health 2022;49(5):707-721. [doi: 10.1007/s10488-022-01194-2] [Medline: 35428931]
34. McMahon FJ. Prediction of treatment outcomes in psychiatry-where do we stand? Dialogues Clin Neurosci 2014 Dec;16(4):455-464. [doi: 10.31887/DCNS.2014.16.4/fmcmahon] [Medline: 25733951]
35. Lorenzo-Luaces L, DeRubeis RJ, van Straten A, Tiemens B. A prognostic index (PI) as a moderator of outcomes in the treatment of depression: a proof of concept combining multiple variables to inform risk-stratified stepped care models. J Affect Disord 2017 Apr 15;213:78-85. [doi: 10.1016/j.jad.2017.02.010] [Medline: 28199892]
36. van Smeden M, Moons KG, de Groot JA, Collins GS, Altman DG, Eijkemans MJ, et al. Sample size for binary logistic prediction models: beyond events per variable criteria. Stat Methods Med Res 2019 Aug;28(8):2455-2474. [doi: 10.1177/0962280218784726] [Medline: 29966490]

Abbreviations

AUC: area under the curve
LASSO: least absolute shrinkage and selection operator
NPV: negative predictive value
OQ-45.2: Outcome Questionnaire
PPV: positive predictive value
ROM: routine outcome monitoring

TRIPOD: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis

Edited by C Lovis; submitted 15.11.22; peer-reviewed by D Meyer, R Bond; revised version received 03.02.23; accepted 24.03.23; published 23.08.23.

Please cite as:
Van Mens K, Lokkerbol J, Wijnen B, Janssen R, de Lange R, Tiemens B
Predicting Undesired Treatment Outcomes With Machine Learning in Mental Health Care: Multisite Study
JMIR Med Inform 2023;11:e44322
URL: https://medinform.jmir.org/2023/1/e44322
doi:10.2196/44322

Social Determinants of Health Documentation in Structured and Unstructured Clinical Data of Patients With Diabetes: Comparative Analysis

Shivani Mehta[1], MPH; Courtney R Lyles[1,2,3,4], PhD; Anna D Rubinsky[5], PhD; Kathryn E Kemper[1], MPH; Judith Auerbach[6], PhD; Urmimala Sarkar[2,3], MD; Laura Gottlieb[7], MD; William Brown III[1,2,4,8,9], PhD, DrPH

1
2
3
4
5
6
7
8
9

Corresponding Author:
Shivani Mehta, MPH

Abstract

Background: Electronic health records (EHRs) have yet to fully capture social determinants of health (SDOH) due to challenges such as nonexistent or inconsistent data capture tools across clinics, lack of time, and the burden of extra steps for the clinician. However, patient clinical notes (unstructured data) may be a better source of patient-related SDOH information.

Objective: It is unclear how accurately EHR data reflect patients' lived experience of SDOH. The manual process of retrieving SDOH information from clinical notes is time-consuming and not feasible. We leveraged two high-throughput tools to identify SDOH mappings to structured and unstructured patient data: PatientExploreR and Electronic Medical Record Search Engine (EMERSE).

Methods: We included adult patients (≥18 years of age) receiving primary care for their diabetes at the University of California, San Francisco (UCSF), from January 1, 2018, to December 31, 2019. We used expert raters to develop a corpus using SDOH in the compendium as a knowledge base as targets for the natural language processing (NLP) text string mapping to find string stems, roots, and syntactic similarities in the clinical notes of patients with diabetes. We applied advanced built-in EMERSE NLP query parsers implemented with JavaCC.

Results: We included 4283 adult patients receiving primary care for diabetes at UCSF. Our study revealed that SDOH may be more significant in the lives of patients with diabetes than is evident from structured data recorded on EHRs. With the application of EMERSE NLP rules, we uncovered additional information from patient clinical notes on problems related to social connectionsisolation, employment, financial insecurity, housing insecurity, food insecurity, education, and stress.

Conclusions: We discovered more patient information related to SDOH in unstructured data than in structured data. The application of this technique and further investment in similar user-friendly tools and infrastructure to extract SDOH information from unstructured data may help to identify the range of social conditions that influence patients' disease experiences and inform clinical decision-making.

(JMIR Med Inform 2023;11:e46159) doi:10.2196/46159

KEYWORDS

natural language processing; diabetes mellitus; medical informatics applications; social determinants of health; NLP; machine learning; diabetes; diabetic; EHR; electronic health record; search engine; free text; unstructured data; text string

Introduction

There is growing recognition that addressing social determinants of health (SDOH)—the conditions in which people are born, grow, work, live, and age—in patient care is necessary for achieving optimal and equitable diabetes outcomes [1,2]. Prior evidence has shown that SDOH, particularly related to low socioeconomic status, affect disparities in the health care

experience of patients with diabetes [3-5]. It is therefore imperative to better understand and intervene on SDOH to prevent negative clinical outcomes for patients and other downstream diabetes health care burdens and disparities [6]. SDOH can impact health equity in both a positive and negative way, thus leading to a gradient of health outcomes [7]. In this study, we focused on SDOH as a social risk factor on health outcomes.

Electronic health records (EHRs) are now becoming a resource to understand patients' SDOH context in ways that could inform clinical practice. However, it remains unclear how accurately EHR data reflect patients' lived experience of SDOH. Historically, EHRs have yet to fully capture SDOH due to challenges such as nonexistent or inconsistent data capture tools across clinics, lack of time and training, the burden of extra steps for the clinician, and the need for manual input, which can be a slow process [8]. Although structured data fields in EHRs for screening SDOH using *International Classification of Diseases (ICD)* codes have become more widespread, these are often not used by clinicians [9]. A better source of SDOH data from the EHR may be unstructured clinical notes, which provide qualitative detail beyond what is captured in structured data fields.

SDOH embedded in clinical notes could be captured quickly using natural language processing (NLP), but a corpus is hard to generate, and data access can be challenging, often requiring advanced programming skills. Novel and innovative high-throughput tools that automate and streamline the process of extracting SDOH data from clinical notes would prove useful to researchers and clinicians without advanced programming skills. Additionally, creating a high-throughput method of identifying SDOH mappings to structured and unstructured patient data has the potential to reduce physician charting burden and improve SDOH data in the EHR.

In 2018, researchers from the University of California, San Francisco (UCSF) created the Compendium of Medical Terminology Codes for Social Risk Factors that maps SDOH to existing ICD codes [10] (referred to hereafter as SDOH ICD Compendium or Compendium). The Compendium contains codes related to 20 SDOH-related risk and resilience factors from four medical vocabularies (LOINC, SNOMED CT, *International Classification of Diseases, Tenth Revision, Clinical Modification [ICD-10-CM]*, and Current Procedural Terminology) [10]. The Compendium allows us to identify existing codes related to social risk factors and their ontology.

In this study, we additionally leveraged two high-throughput tools for identifying SDOH mappings to structured (ICD codes) and unstructured (clinical notes) patient data: PatientExploreR and Electronic Medical Record Search Engine (EMERSE) [11]. We used these existing tools to first identify a cohort of patients within the EHR and explore the structured SDOH ICD Compendium codes (within PatientExploreR) and then to explore textual/unstructured data within the notes of the same patient population using the EMERSE NLP platform (grounded in the terminology from the SDOH ICD Compendium). This allowed us to identify and compare SDOH documentation in both structured and unstructured EHR data in records from patients with diabetes. Our working hypothesis was that these tools would reveal greater information about SDOH among these patients—through the mining of unstructured data—than is captured solely by structured data.

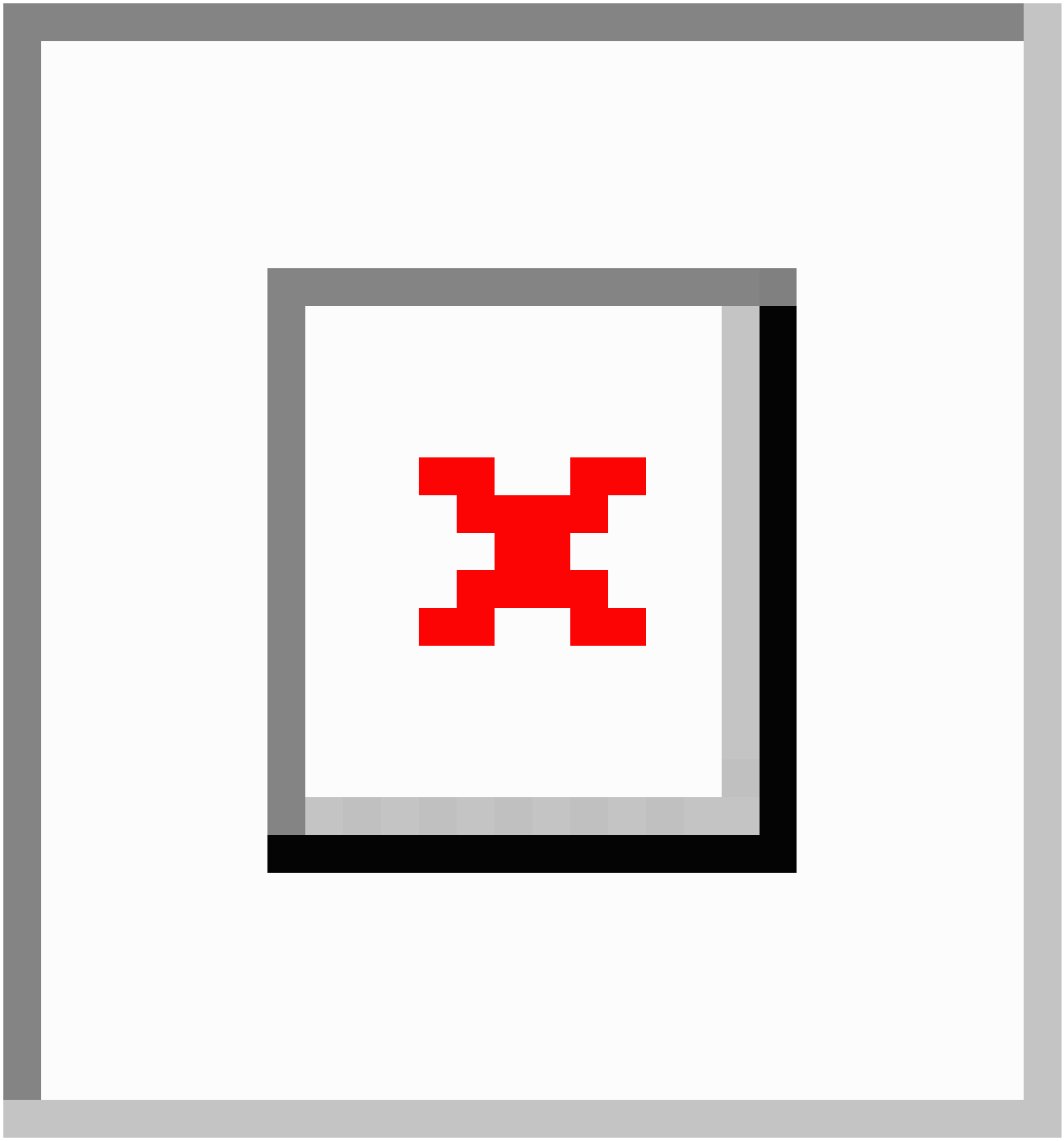
Methods

Deidentified Clinical Data Warehouse and PatientExploreR

UCSF EHR data were extracted using the SQL-based deidentified Clinical Data Warehouse (De-ID CDW) and PatientExploreR. The De-ID CDW is a deidentified database copy of high-value UCSF EHR data (Figure 1). De-ID CDW files are updated monthly, are not subject to Health Insurance Portability and Accountability Act (HIPAA) restrictions on research use, do not require institutional review board approval or an honest broker intermediary, and are available to the UCSF research community at no charge.

The De-ID CDW includes data from UCSF's Epic-based EHR tool and historical EHR data prior to Epic adoption. The De-ID CDW includes the following data elements from the Epic EHR at UCSF Health: patient demographic and geographic information, allergies, billing, coverage, diagnoses, encounters, immunizations, lab, medication orders, procedure orders, providers, clinical notes, and vitals.

Figure 1. Clinical data landscape. CDW: Clinical Data Warehouse; DB: database; De-ID: deidentified; OMOP: Observational Medical Outcomes Partnership; PACS: picture archiving and communication system; Pat Exp: PatientExploreR.



PatientExploreR is a user-friendly R Shiny application that enables rule-based mining of structured, clinical, patient-level interactive dynamic reports using Boolean operators and provides auto-generated visualization of clinical data. PatientExploreR's data pipeline comes from the De-ID CDW, and exploration of the EHR data requires no advanced programming skills, as PatientExploreR data can be queried and extracted in a web-based format.

Data Inclusion Criteria

For this study, we queried PatientExploreR to identify all adult patients (≥18 years of age) receiving primary care services for diabetes between January 1, 2018, to December 31, 2019. Primary care patients were defined as those who had completed two office visits with a primary care department on different dates of service and who had a documented encounter diagnosis of diabetes (type I or type II; *ICD-10-CM*: E10, E11) [12]. We excluded patients who were receiving only specialist care for diabetes as there may be systematic differences in patients receiving specialist care compared to primary care. Additionally, a specialist's documentation related to SDOH may differ from that of primary care physicians and be less generalizable [13].

EMERSE NLP Methods

EMERSE clinical notes are deidentified through automated machine redaction using a protected health information filter [14] (Figure 2). A visualization of the machine-redacted clinical notes data flow is provided in Figure 3. We used EMERSE to

extract clinical notes through a user-friendly interface [11]. We applied advanced built-in NLP query parsers implemented with JavaCC. The Lucene package enabled us to create our own rule-based approach queries through an application programming interface and provided parsing, tokenization features, and proximity searches [15]. We included clinical notes categorized as progress notes, telephone encounters, history and physical examinations, and assessment and plan notes.

Figure 2. Process of Philter deidentification software for University of California, San Francisco, clinical notes. DB: database; HIPAA: Health Insurance Portability and Accountability Act.

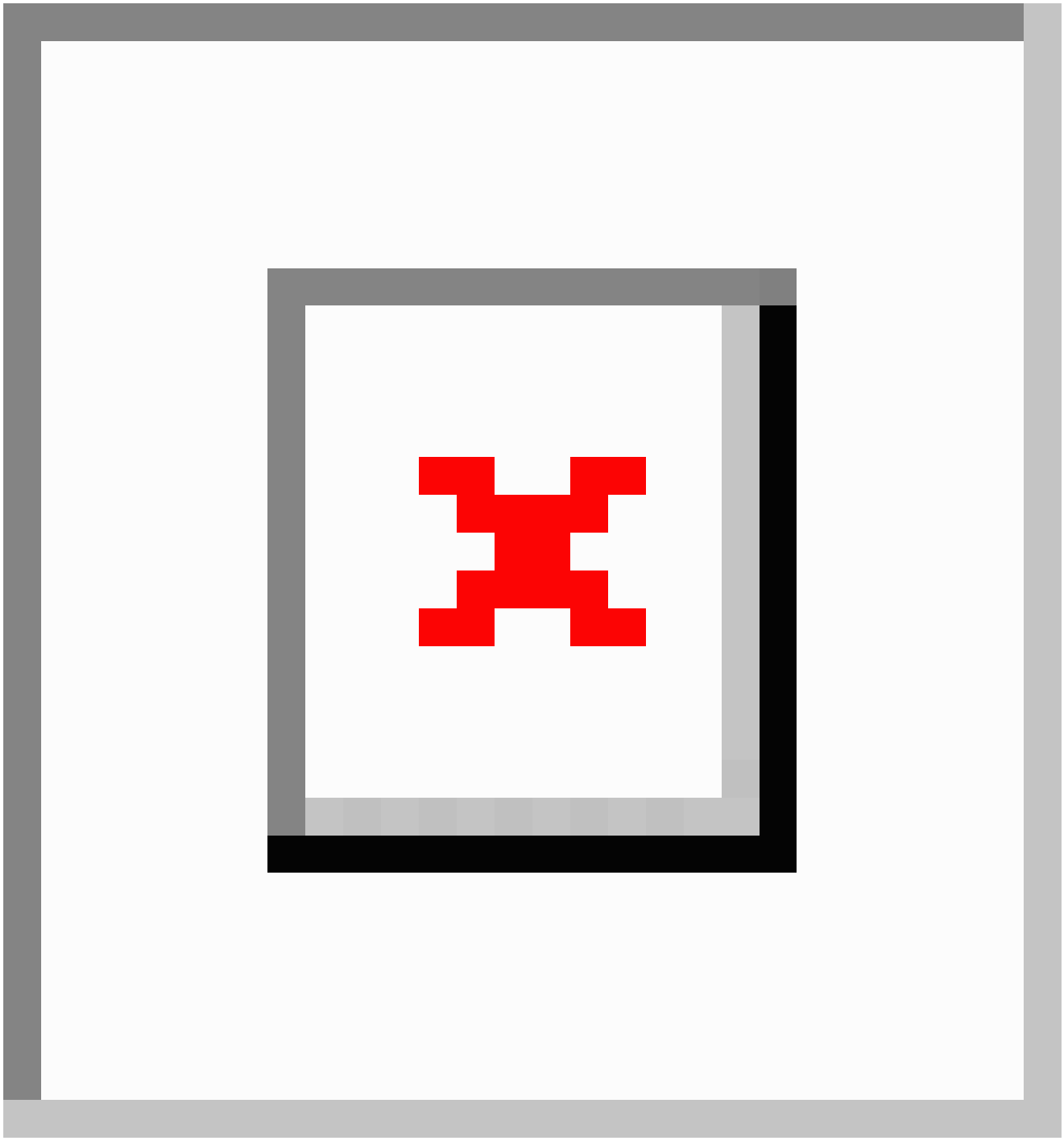
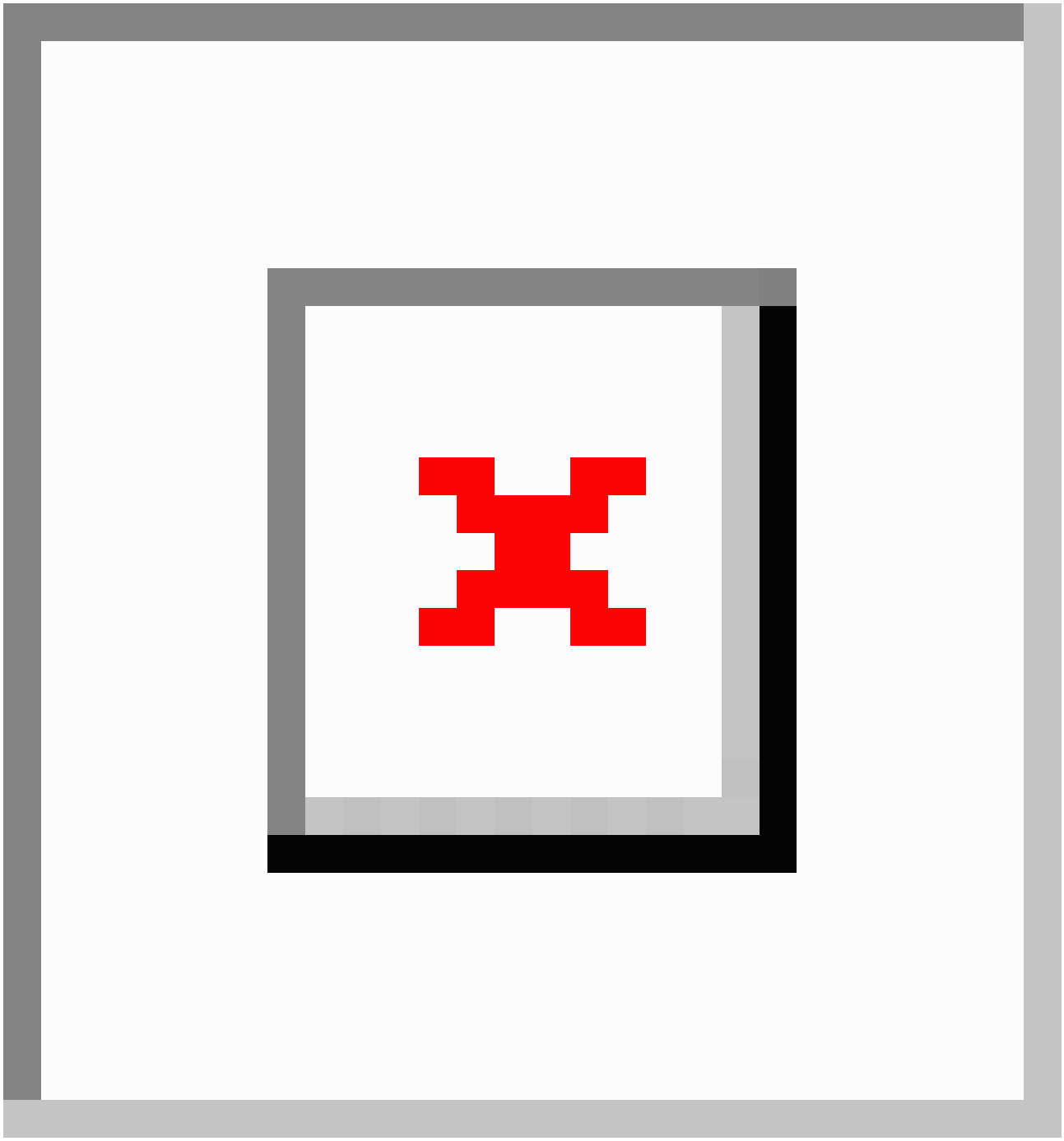


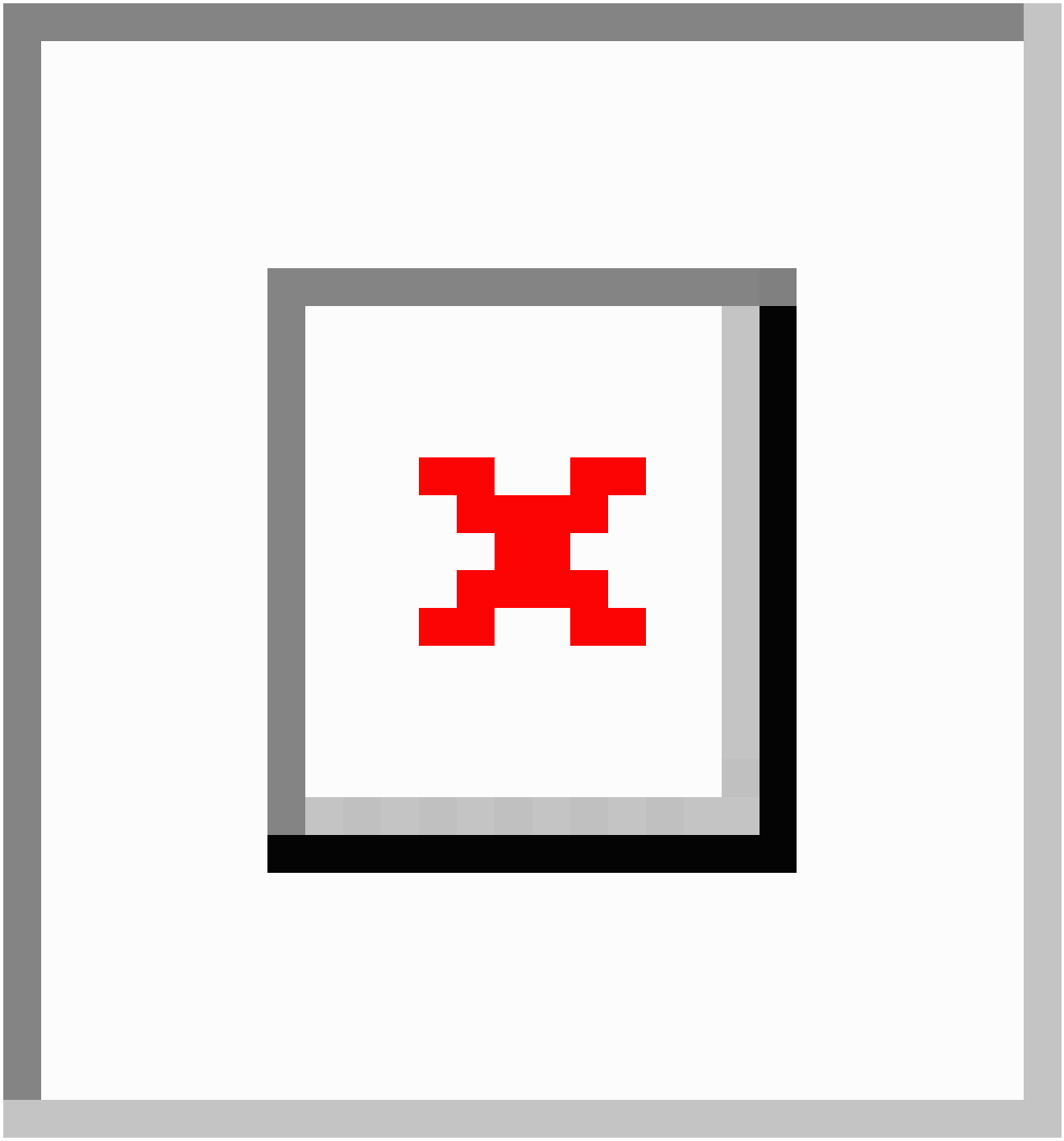
Figure 3. Schematic illustrating machine-redacted clinical notes data flow. SCP: secure copy protocol.



Text Corpus for NLP

For the same cohort of patients with diabetes identified within PatientExploreR, we then conducted a second exploration of SDOH documentation within the unstructured clinical notes (Figure 4). We linked the deidentified patient identifiers from PatientExploreR to the EMERSE platform, in which we were able to explore retracted clinical notes for the same patients during their primary care encounters within the same time period.

Figure 4. Study flow. CDW: Clinical Data Warehouse; DeID: deidentified; EMERSE: Electronic Medical Record Search Engine; MRN: medical record number; NLP: natural language processing.



To compare the SDOH ICD Compendium codes from the structured EHR data to the unstructured clinical notes, we first transformed the Compendium into a set of textual search terms and concepts that would emerge within the free-text sections of the physician's note. We used expert raters to develop a corpus using SDOH in the Compendium as a knowledge base as targets for the NLP text string mapping to find string stems, roots, and syntactic similarities in the clinical notes of patients with diabetes. To reduce false positives, we applied Boolean logic and proximity searches to create an exclusion term list within the NLP rule (Table 1). We determined a priori that the threshold to stop making changes to a rule was when it impacted less than 1% of the cohort.

Table . Natural language processing (NLP) rules.

SDOH[a]	NLP rule
Social connections/isolation	(“social isolation”~10th OR “socially isolated”~10th OR “feeling lonely”~10th OR “Loneliness” OR “isolation sad”~5) NOT (“no loneliness”~5 OR “not lonely”~5 OR “no social isolation” OR “denies loneliness”~4 OR “doesn't feel lonely” OR “isolation 14”~6 OR “can lead to loneliness and isolation” OR “isolation quarantine”~10th OR “isolation none”~5)
Employment	(unemploy* OR “job loss” OR “job fired”~10th OR “job worry”~10th OR “job ruined”~5 OR “job issues”~10th OR “problems at work” OR jobless* OR “does not get hired” OR “looking for work” OR “work fired”~10th OR “stressors work”~4 OR “out of work”)
Housing	((homeless* OR “housing instability” OR “unstable housing” OR evict* OR “shelter” OR “mold house”~10th OR “stressful home situation” OR “search for new place”) OR (“home safety” AND “home environment”)) NOT (“Homeless clients only” OR “volunteering homeless”~10th OR “would homeless”~5 OR “work homeless”~10th OR “homeless mother”~10th OR “homeless father”~10th OR “shelter in place” OR “shelter at home” OR “face tent” OR “oxygen tent”)
Food	(“food insecurity” OR “food insecure” OR “food pantry” OR “food stamp”) NOT (“FOOD INSECURITY: Negative” OR “Denies food insecurity”~10th OR “Food insecurity - worry” OR “Food secure” OR “No food insecurity” OR “Food insecure?” OR “No concerns raised re: food insecurity” OR “does not food pantry”~10)
Education	(Illitera* OR “lack of education” OR “poor education” OR “cannot read” OR “unable to read”) NOT (“label ripped”~3 OR “glucometer” OR “eyesight” OR “vision” OR “small print”)
Finance	(“Poverty” OR “low income” OR “no income” OR “financial difficulty” OR “financial difficulty” OR “financial difficulties” OR “financial issues” OR “financial burden” OR “financial assistance” OR “financial strain” OR “financial support” OR “financial need”) NOT (“not on file”, “if you qualify” OR “none” OR “doesn't qualify”~5 OR “resources”)
Stress	(“family stress”~5 OR "stressed” OR “stressful life”~5 OR “emotional stress” OR “headache stress”~5 OR “feels stressed”~5 OR “very stressed” OR “life stress”) NOT (score OR lab OR echo OR fracture OR myocardial OR perfusion OR exercise OR ecg OR test OR myocardial OR calculate OR ischemia OR ulcer OR induce OR “stressed importance” OR “stressed good”)

[a]SDOH: social determinants of health.

Validation

Two independent reviewers (SM and JA) manually assessed the clinical notes for each SDOH domain to validate the classification performance of the NLP rule. One reviewer (SM) manually reviewed 100% of the clinical notes from each SDOH domain to validate the NLP rule’s classification performance. Reviewers tagged the clinical note as a true positive if the narrative had at least one mention of the social risk factor associated with the respective SDOH domain (Table 2).

Table . Examples of true-positive and false-positive terminologies.

SDOH[a] domain	True-positive terminology	False-positive terminology
Social connections/ isolation	"Pt has difficulties with her mother and social isolation"	"She is deeply concerned about her son's social isolation"
Employment	"Currently unemployed"	"Unemployed son"
Housing	"Pt with unstable housing situation, homeless"	"Volunteers at the animal shelter"
Food insecurity	"Diet remains problem with severe food insecurity"	"Food insecurity: none"
Education	"Never went to school and cannot read"	"Cannot read fine print as well, notes older glasses work better at near"
Finance	"Pt has low income subsidy, financial difficulties"	"Withdrawn cognition: poverty of thought"
Stress	"Feeling very stressed"	"Wife is stressed or complaining"

[a]SDOH: social determinants of health.

A second reviewer (JA) conducted a 2-step validation process. First, all the patients with clinical notes tagged "positive" for SDOH social risk factors by SM were reviewed by JA to ascertain the observed proportional agreement. Second, we randomly sampled 10% of clinical notes from each SDOH domain to ascertain the interrater agreement between SM and JA.

Statistical Analysis

We use the Center for Medicaid Services *ICD-10* Z code groupings to calculate the prevalence of patients with SDOH documented within their structured data [16]. Our first goal was to understand SDOH documentation discrepancies between structured and unstructured clinical notes. Our second goal was to use a user-friendly informatics tool, EMERSE, to develop an NLP rule for each SDOH domain that was able to identify patients with clinical notes containing documentation of the SDOH domains. As part of the validation process, we calculated the proportion of observed agreement and Cohen kappa between the two independent reviewers for each SDOH domain (Multimedia Appendix 1).

Ethical Considerations

The institutional review board at the University of California San Francisco approved this study (IRB number: 18-25696).

Results

We identified 4283 adult (≥18 years of age) patients with 30,288 clinical notes receiving primary care for their diabetes (type I or type II; *ICD-10-CM*: E10, E11) at UCSF from January 1, 2018, to December 31, 2019. In the structured data, 16 (0.38%) patients had *ICD-10* Z60 codes for social connectionsisolation, 14 (0.33%) patients for stress, 4 (0.09%) patients for employment insecurity, 26 (0.61%) patients for housing insecurity, 39 (0.91%) patients for food insecurity, 4 (0.09%) patients for problems related to education, and 4 (0.09%) patients for financial insecurity (Table 3).

Table . Prevalence of patients with SDOH documentation in structured and unstructured data.

SDOH[a] (*ICD-10*[b] code)	Patients in structured data (n=4283), n (%)	Patients in unstructured data (n=4283), n (%)
Social connections/isolation (60.2, 60.4, 60.8)	16 (0.38)	197 (4.60)
Employment (56.0, 56.1, 56.2, 56.89, 56.9)	4 (0.09)	197 (4.60)
Housing (59.0, 59.1, 59.8)	26 (0.61)	111 (2.59)
Food (59.4, 59.41)	39 (0.91)	102 (2.38)
Education (55.0, 55.1, 55.2, 55.3, 55.4, 55.8, 55.9)	4 (0.09)	35 (0.82)
Finance (59.5, 59.6, 59.7)	4 (0.09)	113 (2.64)
Stress (63.7, 63.79, 73.2, 73.3)	14 (0.33)	222 (5.18)

[a]SDOH: social determinants of health.

[b] *ICD-10: International Classification of Diseases, Tenth Revision*

Social Connections/Isolation

Social connections/isolation (*ICD-10-CM* Z60) was defined as a lack of social connections or feelings of isolation or loneliness [10,16]. The NLP rule identified a total of 313 patients with documentation of social connections/isolation in their clinical notes. Of the 313 patients, 15 had a confirmed *ICD-10-CM* Z60 groupings diagnosis within their structured data, and 298 patients did not. A manual review of the clinical notes confirmed social connections/isolation problems for 197 (62.9%) of the 313 patients (Table 4).

Table . Manual review of clinical notes identified by the EMERSE NLP rules.

EMERSE[a] NLP[b] rule (+)	Manual review (+), n	Manual review (–), n	Total, n
Social connections/isolation	197	116	313
Employment	197	161	358
Housing	111	316	427
Food	102	55	157
Education	35	36	71
Finance	113	98	211
Stress	222	288	510

[a]EMERSE: Electronic Medical Record Search Engine.

[b]NLP: natural language processing.

Employment Security

Employment insecurity (*ICD-10-CM* Z56) was defined as problems related to employment, unemployment, job loss, and work-related stressors [10,16]. The NLP rule identified a total of 358 patients with documentation of employment insecurity in their clinical notes. Of the 358 patients, 3 had a confirmed *ICD-10-CM* Z56 diagnosis and 355 patients did not. One patient did not have any clinical notes registered in the EMERSE system. Among 358 patients identified by the NLP rule, a manual review of the clinical notes confirmed problems related to employment for 197 (55%) patients (Table 4).

Housing Security and Quality

This category included homelessness, problems with eviction, unsafe housing conditions (eg, mold), and unstable housing using the *ICD-10-CM* Z59 groupings [16]. The EMERSE NLP rule identified a total of 448 patients with documentation of housing insecurity/poor quality in their clinical notes. Of the 448 patients, 23 had confirmed Z59 diagnosis in their structured data and 425 patients did not. Among the 448 patients identified by the NLP rule, a manual review of the clinical notes confirmed problems with housing security and quality for 111 (24.8%) patients (Table 4).

Food Security

Food insecurity (*ICD-10-CM* Z59.4) was defined as a lack of adequate food or intermittent access to food [10,16]. The NLP rule identified a total of 157 patients with documentation of food insecurity in their clinical notes. Of the 157 patients, 39 had a confirmed *ICD-10-CM* Z59.4 or *ICD-10-CM* Z59.41 diagnosis code in their structured data and 118 patients did not. Among 118 patients identified by the NLP rule, a manual review of the clinical notes confirmed food insecurity for 102 (65%) patients (Table 4).

Education

The education category included patients with problems related to education, unable to read/write, or no formal education using the *ICD-10-CM* Z55 grouping [16]. The NLP rule identified a total of 71 patients with documentation of problems related to education in their clinical notes. Of the 71 patients, 4 had a confirmed *ICD-10* Z55 diagnosis code in their structured data and 67 did not. Among the 71 patients identified by the NLP rule, a manual review of the clinical notes confirmed problems related to education for 35 (49.3%) patients (Table 4).

Finance

Financial insecurity (*ICD-10* Z59.5) was defined as patients reporting financial burdens, low income, poverty, or no income [10,16]. The NLP identified a total of 211 patients with documentation of financial insecurity in their clinical notes. Of the 211 patients, 4 had a confirmed *ICD-10* Z59.5, *ICD-10* Z59.6, or *ICD-10* Z59.7 diagnosis code in their structured data and 207 did not. Among the 211 patients identified by the NLP rule, a manual review of the clinical notes confirmed financial insecurity for 113 (53.6%) patients (Table 4).

Stress

Stress was defined as the lack of relaxation and leisure, and difficulties with life management [10,16]. The NLP rule identified a total of 510 patients with documentation of stress in their clinical notes. Of the 510 patients, 11 had a confirmed *ICD-10* Z63.7, *ICD-10* Z63.79, *ICD-10* Z73.2, or *ICD-10* Z73.3 diagnosis code in their structured data and 499 did not. Among the 510 patients identified by the NLP rule, a manual review of the clinical notes confirmed stress for 222 (43.5%) patients (Table 4).

Interrater Reliability

Observed proportional agreement between both reviewers ranged between 0.98 to 1 for the SDOH domains. The observed proportional agreement refers to the clinical notes in which both reviewers one and two have flagged as a positive for a social risk factor. Cohen kappa ranged from 0.21 to 1 (Multimedia Appendix 1). The validation process allowed us to understand the performance of the NLP rule's classification. Overall, we discovered how much more the unstructured data yields about a patient's SDOH in comparison to structured data.

Discussion

Findings

We included seven SDOH domains—social connections/isolation, problems related to employment, financial insecurity, housing insecurity, food insecurity, education, and stress—and conducted a manual review of clinical notes to validate the SDOH identification. Our study identified a greater

proportion of individuals with diabetes who have an SDOH documented in their EHR when including clinical note data instead of structured data fields alone. In our sample, clinicians frequently captured information in their clinical notes about SDOH in the daily lives of their patients with diabetes, but they did not transfer it to the structured data field on the record, which is a core implementation consideration as the federal government and other agencies are looking to incentivize SDOH screening in the near future [17]. These documentation gaps may contribute to an underestimation of the overall impact of social (including material) and psychological factors on the health outcomes of people with diabetes that contribute to ongoing health disparities.

To the extent that information about SDOH is already being captured by clinicians in unstructured fields, informatics tools like NLP might be used to decrease new clinician structured field documentation burdens. The identification and classification of patients with SDOH using NLP methods is a complex process that involves the understanding of clinical note semantics, lexicon development, categorization, and manual validation.

There was a wide variation in the prevalence of SDOH elements in the unstructured data versus the structured data. The range of variation in the unstructured data depended on the SDOH domain—from 111 (24.8%) patients for housing insecurity compared to 197 (62.9%) patients for social connections/isolation. The findings highlight that future descriptive research should combine the usage of structured and unstructured data.

Comparability

Our study findings are consistent with prior studies that found that EHR structured data underestimates SDOH. These studies found that less than 1% of cohorts had respective *ICD-10-CM* diagnosis codes for SDOH [18-21] documentation. Previous studies have shown that documentation about SDOH such as housing insecurity or lack of social connections or isolation, is 2-fold higher in unstructured data than in structured data [21-24]. However, none of these studies focused on patients with chronic health conditions like diabetes.

Strengths and Limitations

To our knowledge, this is the first study to use PatientExploreR and EMERSE, two high-throughput tools, to identify SDOH mapping in structured and unstructured data for a population of patients with a specific chronic health condition. Neither tool requires users to have prior expertise in programming skills, which makes it accessible to a wider audience of clinicians and researchers. We used the compendium of medical terminology codes for social risk factors as a new data source to generate a corpus. The wider application of this adaptable technique may help to more robustly identify social factors that influence disease management and outcomes for a range of diseases and conditions and inform clinical decision-making.

There are several limitations of this study. This study was conducted using patient-level data from the UCSF medical system, which may limit external validity to the general population of patients with diabetes [25]. However, future work includes validating our NLP rules for a different patient population within the UCSF medical system. It is important to note that our inclusion criteria required patients to have a diagnosis code for diabetes, and this may have missed patients who had diabetes detected via medications or laboratory testing. Although we validated the NLP rule classification performance by manually reviewing the clinical notes that EMERSE deemed as containing SDOH documentation, we were unable to manually validate the clinical notes that our NLP rule did not pick up. This is a limitation as we were unable to calculate sensitivity, specificity, and common metrics to understand our NLP rule's performance. Our NLP rules did not perform well for the following SDOH domains as we identified more false positives than true positives: housing security and quality, financial insecurity, and stress. This warrants further optimization to understand how these SDOHs are characterized within the clinical notes. Some SDOH domains may be more nuanced in terms of the language providers use to note them. This study discovered many false positives from the cases identified by the EMERSE NLP rule. High rates of false positives warrant further optimization of our NLP rule and understanding semantic differences of how SDOH are characterized within patient clinical notes. Future work will focus on enhancing the NLP rule and significant curation. Given the descriptive nature of this study, we did not assess the effects of the temporality of a patient's SDOH, but we conducted a chart review to validate and assess the patient history of SDOH to the extent possible in unstructured notes.

Despite these limitations, this method has proven useful for clinicians and researchers interested in high-throughput ways to capture additional SDOH information related to patients to inform clinical decision-making. The ability to identify patients who are at risk via a streamlined high-throughput method can prevent downstream health burdens of social risk factors. Future work could focus on developing a rule-based machine learning algorithm to create and refine NLP rules associated with the other SDOH domains (eg, inadequate access to health care, incarceration, safety, and transportation barriers). Additionally, it is important for future work to understand the semantic variations that are used to characterize SDOH in clinical notes. Future research in this area to understand whether the performance of the NLP rule differed by certain patient characteristics (eg, age, race, and sex) would be valuable.

Conclusion

Using unstructured data of patients with diabetes via EMERSE, we discovered more patient information related to a set of SDOH than we identified using structured data alone. Application of this technique, and future investments in similar user-friendly tools and infrastructure for capturing information from unstructured EHR data, may help to identify the range of social conditions that influence patients' disease experience and inform clinical decision-making. If these data lead to improvements in clinical care and connections to social services, they are likely to result in improved patient health outcomes and, ideally, contribute to reducing health disparities.

Acknowledgments

SM is funded by a grant from the National Institute on Minority Health and Health Disparities of the National Institutes of Health under award T32MD015070. ADR, CRL, KEK, US, and WB are supported by the National Library of Medicine (R01LM013045). WB is supported by the National Center for Advancing Translational Sciences (KL2TR001870), the National Institute on Drug Abuse of the National Institutes of Health (K01DA055081), and the Agency for Healthcare Research and Quality (K12HS026383). JA is funded by a grant from the National Institute on Minority Health and Health Disparities and Health Resources and Services Administration. LG is funded by a grant from the National Institute on Minority Health and Health Disparities and the Robert Wood Johnson Foundation.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Contingency table for interrater agreement per social determinant of health domain.
[DOCX File, 17 KB - medinform_v11i1e46159_app1.docx]

References

1. Andermann A, CLEAR Collaboration. Taking action on the social determinants of health in clinical practice: a framework for health professionals. CMAJ 2016 Dec 6;188(17-18):E474-E483. [doi: 10.1503/cmaj.160177] [Medline: 27503870]
2. Hill J, Nielsen M, Fox MH. Understanding the social factors that contribute to diabetes: a means to informing health care and social policies for the chronically ill. Perm J 2013;17(2):67-72. [doi: 10.7812/TPP/12-099] [Medline: 23704847]
3. Hill-Briggs F, Adler NE, Berkowitz SA, Chin MH, Gary-Webb TL, Navas-Acien A, et al. Social determinants of health and diabetes: a scientific review. Diabetes Care ;44(1):258-279 dci200053. [doi: 10.2337/dci20-0053] [Medline: 33139407]
4. Ogunwole SM, Golden SH. Social determinants of health and structural inequities-root causes of diabetes disparities. Diabetes Care 2021 Jan;44(1):11-13. [doi: 10.2337/dci20-0060] [Medline: 33571949]
5. Scott A, Chambers D, Goyder E, O'Cathain A. Socioeconomic inequalities in mortality, morbidity and diabetes management for adults with type 1 diabetes: a systematic review. PLoS One 2017 May 10;12(5):e0177210. [doi: 10.1371/journal.pone.0177210] [Medline: 28489876]
6. Walker RJ, Smalls BL, Campbell JA, Strom Williams JL, Egede LE. Impact of social determinants of health on outcomes for type 2 diabetes: a systematic review. Endocrine 2014 Sep;47(1):29-48. [doi: 10.1007/s12020-014-0195-0] [Medline: 24532079]
7. Braveman P, Gottlieb L. The social determinants of health: it's time to consider the causes of the causes. Public Health Rep 2014 ;129(Suppl 2):19-31. [doi: 10.1177/00333549141291S206] [Medline: 24385661]
8. Gold R, Cottrell E, Bunce A, Middendorf M, Hollombe C, Cowburn S, et al. Developing electronic health record (EHR) strategies related to health center patients' social determinants of health. J Am Board Fam Med 2017 ;30(4):428-447. [doi: 10.3122/jabfm.2017.04.170046] [Medline: 28720625]
9. Wang M, Pantell MS, Gottlieb LM, Adler-Milstein J. Documentation and review of social determinants of health data in the EHR: measures and associated insights. J Am Med Inform Assoc 2021 Nov 25;28(12):2608-2616. [doi: 10.1093/jamia/ocab194] [Medline: 34549294]
10. Arons A, DeSilvey S, Fichtenberg C, Gottlieb L, Social Interventions Research & Evaluation Network. Compendium of medical terminology codes for social risk factors. 2018. URL: sirenetwork.ucsf.edu/tools-resources/resources/compendium-medical-terminology-codes-social-risk-factors [accessed 2023-07-4]
11. Hanauer DA, Mei Q, Law J, Khanna R, Zheng K. Supporting information retrieval from electronic health records: a report of University of Michigan's nine-year experience in developing and using the electronic medical record search engine (EMERSE). J Biomed Inform 2015 Jun;55:290-300. [doi: 10.1016/j.jbi.2015.05.003] [Medline: 25979153]
12. O'Neill A, National Committee for Quality Assurance. Comprehensive diabetes care. URL: www.ncqa.org/hedis/measures/comprehensive-diabetes-care/ [accessed 2023-06-28]
13. Pollard SE, Neri PM, Wilcox AR, Volk LA, Williams DH, Schiff GD, et al. How physicians document outpatient visit notes in an electronic health record. Int J Med Inform 2013 Jan;82(1):39-46. [doi: 10.1016/j.ijmedinf.2012.04.002] [Medline: 22542717]
14. Norgeot B, Muenzen K, Peterson TA, Fan X, Glicksberg BS, Schenk G, et al. Protected health information filter (Philter): accurately and securely de-identifying free-text clinical notes. NPJ Digit Med 2020 Apr 14;3(1):1-8. [doi: 10.1038/s41746-020-0258-y] [Medline: 32337372]
15. Apache Software Foundation, Apache Lucene. Package org.apache.lucene.queryparser.classic. URL: lucene.apache.org/core/7_2_1/queryparser/org/apache/lucene/queryparser/classic/package-summary.html#Overview [accessed 2023-06-29]

16. Maksut JL, Hodge C, Van CD, Razmi A, Khau MT, Centers for Medicare & Medicaid Services. Utilization of Z codes for social determinants of health among medicare fee-for-service beneficiaries, 2019. 2021. URL: www.cms.gov/files/document/z-codes-data-highlight.pdf [accessed 2023-07-4]
17. Jacobs DB, Schreiber M, Seshamani M, Tsai D, Fowler E, Fleisher LA. Aligning quality measures across CMS — the universal foundation. N Engl J Med ;388(9):776-779. [doi: 10.1056/NEJMp2215539] [Medline: 36724323]
18. Torres JM, Lawlor J, Colvin JD, Sills MR, Bettenhausen JL, Davidson A, et al. ICD social codes: an underutilized resource for tracking social needs. Med Care 2017 Sep;55(9):810-816. [doi: 10.1097/MLR.0000000000000764] [Medline: 28671930]
19. Kharrazi H, Anzaldi LJ, Hernandez L, Davison A, Boyd CM, Leff B, et al. The value of unstructured electronic health record data in geriatric syndrome case identification. J Am Geriatr Soc 2018 Aug;66(8):1499-1507. [doi: 10.1111/jgs.15411] [Medline: 29972595]
20. Chen T, Dredze M, Weiner JP, Hernandez L, Kimura J, Kharrazi H. Extraction of geriatric syndromes from electronic health record clinical notes: assessment of statistical natural language processing methods. JMIR Med Inform 2019 Mar 26;7(1). [doi: 10.2196/13039] [Medline: 30862607]
21. Gundlapalli AV, Carter ME, Palmer M, Ginter T, Redd A, Pickard S, et al. Using natural language processing on the free text of clinical documents to screen for evidence of homelessness among US veterans. AMIA Annu Symp Proc 2013 Nov 16;2013:537-546. [Medline: 24551356]
22. Bucher BT, Shi J, Pettit RJ, Ferraro J, Chapman WW, Gundlapalli A. Determination of marital status of patients from structured and unstructured electronic healthcare data. AMIA Annu Symp Proc 2020 Mar 4;2019:267-274. [Medline: 32308819]]
23. Navathe AS, Zhong F, Lei VJ, Chang FY, Sordo M, Topaz M, et al. Hospital readmission and social risk factors identified from physician notes. Health Serv Res 2018 Apr;53(2):1110-1136. [doi: 10.1111/1475-6773.12670] [Medline: 28295260]
24. Hatef E, Rouhizadeh M, Tia I, Lasser E, Hill-Briggs F, Marsteller J, et al. Assessing the availability of data on social and behavioral determinants in structured and unstructured electronic health records: a retrospective analysis of a multilevel health care system. JMIR Med Inform 2019 Aug 2;7(3). [doi: 10.2196/13802] [Medline: 31376277]
25. Ceriello A, Barkai L, Christiansen JS, Czupryniak L, Gomis R, Harno K, et al. Diabetes as a case study of chronic disease management with a personalized approach: the role of a structured feedback loop. Diabetes Res Clin Pract 2012 Oct 1;98(1):5-10. [doi: 10.1016/j.diabres.2012.07.005] [Medline: 22917639]

Abbreviations

DeID-CDW: deidentified Clinical Data Warehouse
EHR: electronic health record
EMERSE: Electronic Medical Record Search Engine
HIPAA: Health Insurance Portability and Accountability Act
***ICD*:** *International Classification of Diseases*
***ICD-10-CM*:** *International Classification of Diseases, Tenth Revision, Clinical Modification*
NLP: natural language processing
SDOH: social determinants of health
UCSF: University of California, San Francisco

Edited by J Hefner; submitted 31.01.23; peer-reviewed by A Hirsch, J Zheng; revised version received 06.05.23; accepted 10.06.23; published 22.08.23.

Please cite as:
Mehta S, Lyles CR, Rubinsky AD, Kemper KE, Auerbach J, Sarkar U, Gottlieb L, Brown III W
Social Determinants of Health Documentation in Structured and Unstructured Clinical Data of Patients With Diabetes: Comparative Analysis
JMIR Med Inform 2023;11:e46159
URL: https://medinform.jmir.org/2023/1/e46159
doi:10.2196/46159

Original Paper

Chinese Clinical Named Entity Recognition From Electronic Medical Records Based on Multisemantic Features by Using Robustly Optimized Bidirectional Encoder Representation From Transformers Pretraining Approach Whole Word Masking and Convolutional Neural Networks: Model Development and Validation

Weijie Wang[1], MS; Xiaoying Li[1], PhD; Huiling Ren[1], MS; Dongping Gao[1], PhD; An Fang[1], MS

Institute of Medical Information and Library, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, China

Corresponding Author:
Huiling Ren, MS
Institute of Medical Information and Library
Chinese Academy of Medical Sciences & Peking Union Medical College
69 Dongdan N St
Beijing, 100005
China
Phone: 86 010 52328911
Email: ren.huiling@imicams.ac.cn

Abstract

Background: Clinical electronic medical records (EMRs) contain important information on patients' anatomy, symptoms, examinations, diagnoses, and medications. Large-scale mining of rich medical information from EMRs will provide notable reference value for medical research. With the complexity of Chinese grammar and blurred boundaries of Chinese words, Chinese clinical named entity recognition (CNER) remains a notable challenge. Follow-up tasks such as medical entity structuring, medical entity standardization, medical entity relationship extraction, and medical knowledge graph construction largely depend on medical named entity recognition effects. A promising CNER result would provide reliable support for building domain knowledge graphs, knowledge bases, and knowledge retrieval systems. Furthermore, it would provide research ideas for scientists and medical decision-making references for doctors and even guide patients on disease and health management. Therefore, obtaining excellent CNER results is essential.

Objective: We aimed to propose a Chinese CNER method to learn semantics-enriched representations for comprehensively enhancing machines to understand deep semantic information of EMRs by using multisemantic features, which makes medical information more readable and understandable.

Methods: First, we used Robustly Optimized Bidirectional Encoder Representation from Transformers Pretraining Approach Whole Word Masking (RoBERTa-wwm) with dynamic fusion and Chinese character features, including 5-stroke code, Zheng code, phonological code, and stroke code, extracted by 1-dimensional convolutional neural networks (CNNs) to obtain fine-grained semantic features of Chinese characters. Subsequently, we converted Chinese characters into square images to obtain Chinese character image features from another modality by using a 2-dimensional CNN. Finally, we input multisemantic features into Bidirectional Long Short-Term Memory with Conditional Random Fields to achieve Chinese CNER. The effectiveness of our model was compared with that of the baseline and existing research models, and the features involved in the model were ablated and analyzed to verify the model's effectiveness.

Results: We collected 1379 Yidu-S4K EMRs containing 23,655 entities in 6 categories and 2007 self-annotated EMRs containing 118,643 entities in 7 categories. The experiments showed that our model outperformed the comparison experiments, with F_1-scores of 89.28% and 84.61% on the Yidu-S4K and self-annotated data sets, respectively. The results of the ablation analysis demonstrated that each feature and method we used could improve the entity recognition ability.

Conclusions: Our proposed CNER method would mine the richer deep semantic information in EMRs by multisemantic embedding using RoBERTa-wwm and CNNs, enhancing the semantic recognition of characters at different granularity levels

and improving the generalization capability of the method by achieving information complementarity among different semantic features, thus making the machine semantically understand EMRs and improving the CNER task accuracy.

(JMIR Med Inform 2023;11:e44597) doi:10.2196/44597

KEYWORDS

Chinese clinical named entity recognition; multisemantic features; image feature; Robustly Optimized Bidirectional Encoder Representation from Transformers Pretraining Approach Whole Word Masking; RoBERTa-wwm; convolutional neural network; CNN

Introduction

Background

Abundant medical data have been accumulated since the development of the hospital information system, among which the electronic medical records (EMRs) contain information closely related to patients' diagnosis and treatment processes [1]. As important records of patients' medical activities, effective extraction and use of the medical information contained in EMRs could provide clinical decision-making support for doctors and realize personalized medical guidance and health management for patients. It could also help biomedical researchers discover the tacit medical knowledge, thus providing ideas for studies of the association between diseases, the relationship between symptoms, the prediction of diseases and therapies, complication prediction, comorbidity analysis, etc. The medical information would be rapidly extracted from the unstructured EMRs through named entity recognition (NER). NER is a basic task of natural language processing, which will lay the foundation for the construction of medical knowledge graphs, medical knowledge bases, and so on by steps such as medical entity structuring, medical entity standardization, and medical entity relationship extraction. It will also provide fundamental support for practical application scenarios such as medical knowledge retrieval systems, clinical decision support systems, clinical event extraction, and so on [2,3].

Clinical NER (CNER) refers to the recognition of entities such as anatomy, disease, symptoms, clinical examination, medication, surgical procedure, and so on from EMRs [4,5]. Chinese CNER is more difficult than English NER for several reasons. First, Chinese words lack space segmentation and have blurred boundaries. Second, the composition of a Chinese entity is complex and may contain various figures, letters, and abbreviations. Third, Chinese grammar is complicated, and the same word may represent different entity types in different contexts. Therefore, Chinese CNER remains a research focus.

Recently, the features of radicals for Chinese characters have been widely used to improve the efficiency of different Chinese natural language processing tasks [6-8]. Chinese characters, known for thousands of years, are highly developed morpheme scripts that are still used worldwide with unique ideology [9]. Chinese characters include single-component and multiple-component characters. A single-component character cannot be divided, for example, "心 (heart)," "手 (hand)," and "口 (mouth)," and so on; whereas a multiple-component character is composed of basic components, accounting for >90% of Chinese characters [10], for example, the radical for "呕 (vomit)" and "吐 (vomit)" is "口 (mouth)," and the radical for "肿 (swelling)" and "胀 (swelling)" is "月 (month)," which refers to meat or organs in ancient times. Chinese characters are divided into associative compound characters, indicative characters, pictographic characters, and picto-phonetic characters based on their characteristics. In addition, Chinese characters are also called square characters, as they are square, and there are 8 structures of Chinese characters that are subdivided based on their intrinsic shape and construction. Therefore, Chinese characters contain rich deep semantic information. Applying radicals, phonological codes, shape structures, and other features would help to improve Chinese CNER accuracy.

The contributions of this study are as follows: (1) using pretrained language model (PLM) Robustly Optimized Bidirectional Encoder Representation from Transformers Pretraining Approach Whole Word Masking (RoBERTa-wwm) with a dynamic fusion transformer layer to obtain the semantic features of Chinese characters; (2) using CNNs for extracting the radicals and picto-phonetic features of Chinese characters through the 5-stroke code, Zheng code, phonological code, and stroke code; (3) converting Chinese characters into square images, extracting Chinese character image features from another modality by CNNs, and deeply capturing the pictographic characteristics of Chinese characters; and (4) improving the semantic recognition ability of the model at different levels of granularity, achieving information complementarity between different semantic features, and improving the effect and generalization ability of the model based on multisemantic features.

Related Works

Medical NER

In recent decades, the medical NER is still a research focus. Medical NER research has 3 main development stages as follows: based on dictionaries and rules, based on statistical machine learning, and based on deep learning.

The dictionary-based [11-13] methods need to construct a domain dictionary in advance to achieve medical NER by matching algorithms. The accuracy of this method is relatively higher. However, it may be affected by the large number, strong specialization, and high complexity of Chinese medical entities. In addition, medical terminologies are updated quickly with the rapid development of the medical field, and the lack of new terminologies will also affect medical NER accuracy. The rule-based [14,15] methods need experts in a particular field to formulate the rule templates based on information such as context grammar and structure. However, the rules are poorly universal in different fields. The methods based on dictionaries and rules [16-18] are poorly generalized, time-consuming, and

objective, as much time and labor are required. Therefore, many scholars have gradually applied methods based on statistical machine learning on medical NER. The commonly used methods include maximum entropy [19], support vector machine [20,21], hidden Markov model [22,23], and conditional random fields (CRF) [24,25]. However, these methods rely on large-scale annotation data sets [26] and manual feature selection [13,27]. Moreover, the quality of the selected features will directly affect the medical NER results.

With the continuous development of deep learning, Cocos et al [28] found that deep learning has advantages over traditional machine learning. It can automatically extract the characteristics of various levels and reduce the subjectivity of artificial feature selection. This thereby improves the result accuracy. The commonly used deep learning models include convolutional neural networks (CNNs) [29], recurrent neural networks [30], long short-term memory (LSTM) [31], Word to Vector (Word2Vec) [32], Bidirectional Encoder Representation from Transformers (BERT) [33], and so on. However, fully extracting the data features by using a neural network alone is challenging. Most scholars took Long LSTM-CRF as the main framework to make up for the medical NER deficiency using a single neural network [34]. The Bidirectional LSTM-CRF (BiLSTM-CRF) [35] model was then developed. This model could better capture contextual information as an important milestone in medical NER and has been widely used in the medical field [36,37]. To improve the ability to capture details and extract features of medical NER models, many studies added Word2Vec with static representation [38], Global Vectors for Word Representation [39] with static representation, Embeddings from Language Models (ELMo) [40,41] with dynamic representation, CNN [42], and attention mechanism [43] to the BiLSTM-CRF model. Some studies [44,45] have shown that the application of the BiLSTM-CRF model combined with the word vector generated by BERT could significantly improve medical NER accuracy. BERT provided a more accurate word representation and achieved better task results than traditional word vector methods. As per the specialty of medicine and the characteristics of Chinese characters involved, the clinical dictionaries, root-level features, parts of speech, radicals, and phonological codes have been added in the BiLSTM-CRF model in some studies [46-51] for improving Chinese CNER performance.

PLMs Technique

PLMs are pretrained on a large-scale corpus to obtain prior semantic knowledge from unlabeled text and improve the effectiveness of different downstream tasks. The word vector generated by a bidirectional language model BERT with stacked transformer substructures contains not only the preliminary information from the corpus training but also the encoded contextual information. Some robust versions of BERT have been constructed since BERT was proposed in 2018. For example, the RoBERTa model [52], which replaces the static (MASK) strategy with a dynamic (MASK) strategy, and the words (MASK) in each sequence dynamically change in different epoch trainings. In addition, the RoBERTa model is retrained with bigger batches and longer sequences, and the next-sentence prediction task, which is not related to the downstream task, is canceled during the pretraining. Compared with the BERT model, the RoBERTa model performs better on multiple natural language processing tasks. However, the character-level RoBERTa model does not fit the Chinese natural language processing, as the different segmentation modes between Chinese and English words suffer a limitation of lacking word information. Then, the word-level RoBERTa-wwm model [53] was proposed based on Chinese characteristics, which greatly improved the text representation ability in Chinese [54].

Methods

Data Collection

The Yidu-S4K data set, shared publicly by YiduCloud, is derived from the Chinese EMRs entity recognition task of the China Conference on Knowledge Graph and Semantic Computing 2019 [55]. It contains 1379 EMRs with 6 entity types, including Disease (medically defined disease and diagnoses made by physicians based on etiology, pathophysiology, pathological classification, and clinical staging); Anatomy (anatomical parts of the body where disease, signs, and symptoms occurred); Laboratory (physical or chemical tests performed by the laboratory department in clinical work); Image (imaging [x-ray, computed tomography, magnetic resonance imaging, positron emission tomography-computed tomography, etc], ultrasound, and electrocardiogram); Medicine (specific chemical substances used for disease treatment); and Operation (treatments focused on surgery such as excision and suturing performed by the physician locally on the patient's body).

Self-annotated EMR data, collected from publicly desensitized Chinese EMR websites [56], contain 2007 EMRs. As per the Terminology of Clinical Medicine issued by the National Health Commission of the People's Republic of China, we used the BIO (B signifies the beginning of an entity, I signifies that the word is inside an entity, and O signifies that the word is just a regular word outside of an entity) tagging method to pretag 7 entity types in the EMRs, including Disease (same definition as the Yidu-S4K data set); Symptoms (abnormal manifestations as perceived by the sensory organs of patients and physicians); Anatomy (same definition as the Yidu-S4K data set); Examination (includes imaging examinations and laboratory tests mentioned in the Yidu-S4K data set); Instrument (apparatus and mechanical equipment for disease prevention, diagnosis, treatment, health care, and rehabilitation); Medicine (the same definition as the Yidu-S4K data set); and Operation (same definition as the Yidu-S4K data set). Subsequently, 4 medical experts manually checked and corrected the tags. The interclass correlation efficient consistency test revealed that we had good annotation quality.

The ratio of the training set to the test set of the EMRs was 7:3. The Yidu-S4K data set was preprovided with 1000 EMRs as the training data sets (1000/1379, 72.52%) and 379 EMRs as the test data sets (379/1379, 27.48%). The self-annotated data set was divided by randomization into 1401 EMRs as the training data sets (1401/2007, 69.81%) and 379 EMRs as the

test data sets (606/2007, 30.19%). Table 1 lists the details of the different types of entities in the 2 EMR data sets.

Table 1. The statistics of different types of entities in 2 electronic medical record data sets

Data sets and entity type	Training set, n	Test set, n
Yidu-S4K		
Disease	4212	1323
Anatomy	8426	3094
Laboratory	1195	590
Image	969	348
Medicine	1822	485
Operation	1029	162
All entities	17,653	6002
Self-annotated		
Disease	9470	4504
Symptoms	26,334	11,065
Anatomy	17,877	7588
Examination	19,664	8746
Instrument	1244	560
Medicine	5314	2566
Operation	2578	1133
All entities	82,481	36,162

Ethical Considerations

Ethics approval was not required because the patient's private information was masked by the website.

Experiments Settings

In this study, all the experiments were conducted by Python [57] and PyTorch [58]. Table 2 shows the experimental parameters. The experiments used *RoBERTa-wwm-ext-large* model pretraining data, optimized parameters using Adam W, dropout to prevent overfitting, the batch size of 32, BiLSTM hidden layer dimension of 768, maximum sequence length of 510, RoBERTa-wwm dimension of 768, semantic feature dimension of 124, and image feature dimension of 128. On 2 Chinese CNER data sets, we used the same parameters.

Table 2. Parameter settings.

Parameter	Value
Dropout	0.5
Epoch	Optimization
Optimization	Adam W
Learning rate	0.0001
Batch size	32
BiLSTM[a] hidden layer	768
Max_len	510
RoBERTa-wwm[b] feature dimension	768
Semantic feature dimension	124
Image feature dimension	128

[a]BiLSTM: Bidirectional Long Short-Term Memory.

[b]RoBERTa-wwm: Robustly Optimized Bidirectional Encoder Representation from Transformers Pretraining Approach Whole Word Masking.

Evaluation Metrics

The experiments used precision, recall, and F_1-score to evaluate the model performance. The formulas for each index are as follows:

$$Precision = TP / (TP + FP) \quad (1)$$

$$Recall = TP / (TP + FN) \quad (2)$$

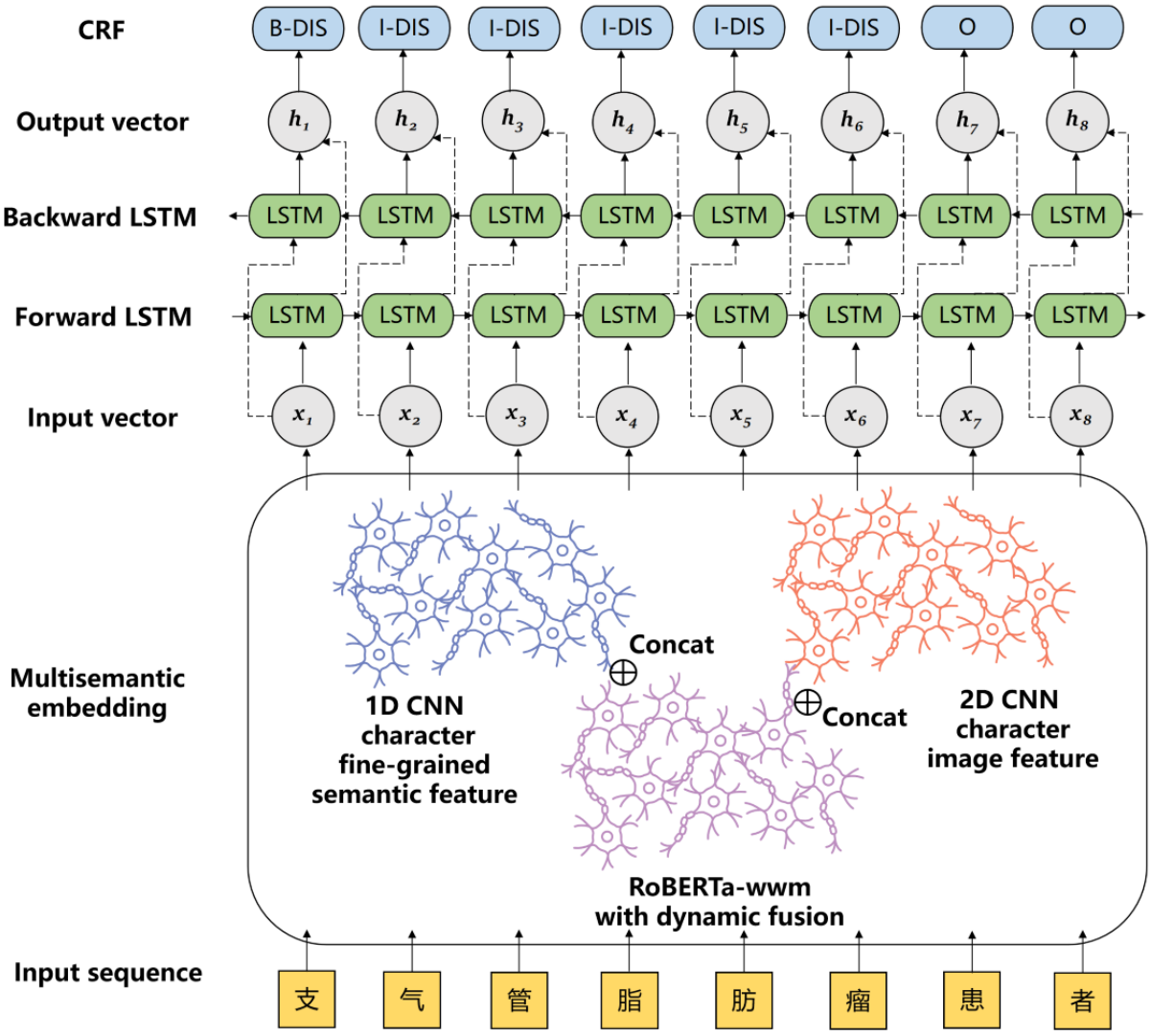
$$F_1\text{-}score = (2 \times precision \times recall) / (precision + recall) \quad (3)$$

where precision is the proportion of positive samples in all samples predicted to be positive; recall is the proportion of positive samples in all positive samples; F_1-score is the harmonic mean of precision and recall; true positive (TP) is the number of positive samples predicted to be positive, that is, the number of correctly recognized entities; false positive (FP) is the number of negative samples predicted to be negative, that is, the number of incorrectly recognized other texts as entities; and false negative (FN) is the number of positive samples predicted to be negative, that is, the number of unrecognized entities.

Model Overview

In this study, we proposed a CNER model based on multisemantic features, as shown in Figure 1. First, we used RoBERTa-wwm, the PLM, to obtain the embedded representation at the word level. Dynamic fusion is performed on the semantic representation generated by each transformer layer to make full use of RoBERTa-wwm representation information. Then, the embedded Chinese character fine-grained feature representation, including the 5-stroke code, Zheng code, phonological code, and stroke code, is extracted by 1D CNN, whereas the embedded Chinese character image representation is extracted from another modality by 2D CNN, with the Chinese characters as square images. Finally, the above multisemantic vectors were input into the BiLSTM layer for encoding and were decoded in the CRF layer to predict the tag probability.

Figure 1. The main architecture of our model. 1D CNN: 1D convolutional neural network; 2D CNN: 2D convolutional neural network; B-DIS: beginning of disease entity; CRF: conditional random fields; h: embedding of output character; I-DIS: inside of disease entity; LSTM: long short-term memory; O: other type; RoBERTa-wwm: Robustly Optimized Bidirectional Encoder Representation from Transformers Pretraining Approach Whole Word Masking; x: embedding of input character.



Multisemantic Embedding Layer

Overview

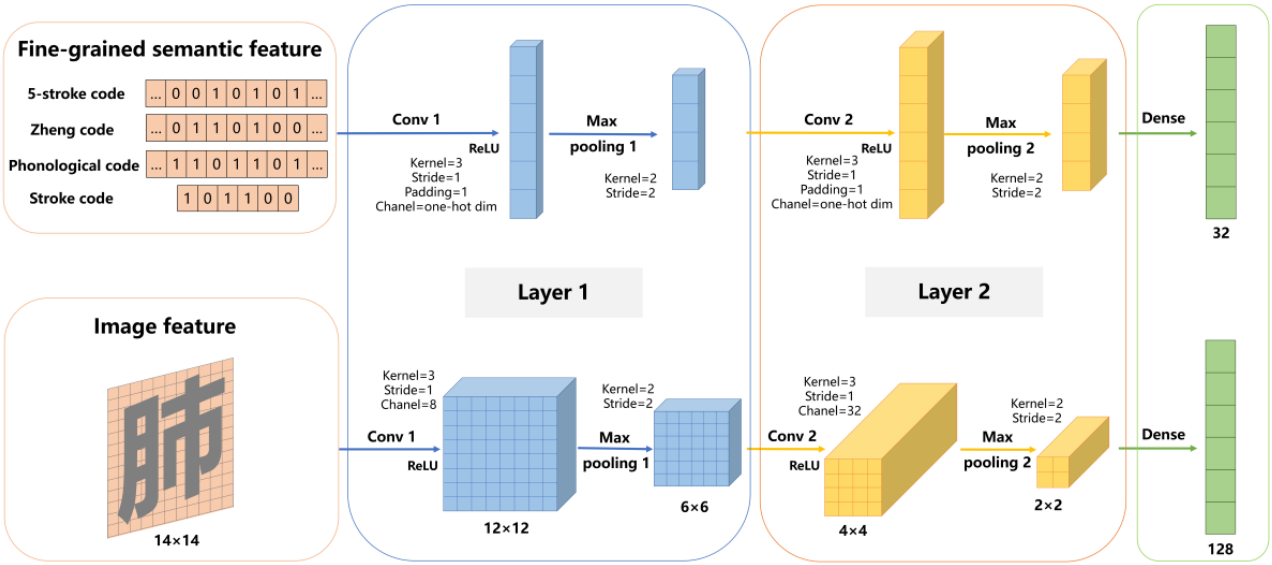
Many Chinese characters have retained their original connotations, as they originated from pictographic characters in ancient times. Moreover, the inherent fine-grained character information contained in Chinese characters often implies more additional semantic information. Accordingly, we obtained the 5-stroke code, Zheng code, phonological code, and stroke code information, as shown in Table 3, of the Chinese characters from ZDIC [59] and embedded them in the model. In addition, Chinese characters are squares, and different shapes and structures express different types of information. Characters with similar intrinsic characteristics may have similar meanings. Therefore, we took Chinese characters as graphics and obtained semantic information on Chinese character connotations from another modality. Multisemantics could obtain information comprehensively and learn a better feature representation by making use of information complementarity and eliminating the redundancy among different semantic features compared with a single-semantic feature, resulting in a more generalized model.

As shown in Figure 2, we converted the Chinese character 5-stroke code, Zheng code, phonological code, and stroke code into one-hot vector encoding and interpreted the Chinese characters as 14×14 images. Subsequently, we used a 2-layer CNN deeply extracting the Chinese character multisemantic features. Through the Convolution layer with the *ReLU* activation function, max pooling layer, and dense layer, we obtained the multisemantic vectors that could be embedded in the BiLSTM layer.

Table 3. Example of Chinese characters' coded information from ZDIC.

Character	5-stroke code	Zheng code	Phonological code	Stroke code
呕 (vomit)	kaqy	jhos	ǒu	2,511,345
吐 (vomit)	kfg	jbvv	tù	251,121
肿 (swelling)	ekhh	qji	zhǒng	35,112,512
胀 (swelling)	etay	qch	zhàng	35,113,154
心 (heart)	nyny	wz	xīn	4544
手 (hand)	rtgh	md	shǒu	3112

Figure 2. The process of obtaining Chinese character multisemantic features by convolutional neural network. ReLU: Rectified Linear Unit function; Conv 1: first convolutional layer; Conv 2: second convolutional layer; Max pooling 1: first max pooling layer; Max pooling 2: second max pooling layer; Dense: dense layer.



RoBERTa-wwm With Dynamic Fusion

When RoBERTa-wwm pretrains the corpus, it is segmented on the language technology platform established by the Harbin Institute of Technology based on Wikipedia content in Chinese, which can provide a basis for achieving wwm. As shown in Figure 3, the word "支气管 (bronchi)" in the RoBERTa-wwm model is completely masked by random wwm, whereas only single characters can be randomly masked in the BERT model, for example, only 1 character "气 (gases)" was masked in the word "支气管 (bronchi)." Thus, the RoBERTa-wwm model can learn the word-level semantic representations in Chinese.

The encoder structure of each transformer layer of the BERT model outputs had different abstract representations of grammar, semantics, and real knowledge in sentences. Studies have confirmed that each layer of the BERT model represents text information differently through 12 natural language processing tasks [60]. As shown in Figure 4, the low transformer mainly learns and encodes surface features; the middle transformer

learns and encodes syntactic features; and the high transformer learns and encodes semantic features.

The transformer structure of the RoBERTa-wwm model is consistent with that of the BERT model. To make full use of the representation information of each transformer layer, we used the RoBERTa-wwm model with dynamic fusion [61]. This helped in assigning the initial weight to the representation vector of 12 transformer layers, determining the weight during training, and weighing the representation vector generated by each layer.

Figure 3. Mask process of Bidirectional Encoder Representation from Transformers (BERT) and Robustly Optimized Bidirectional Encoder Representation from Transformers Pretraining Approach Whole Word Masking (RoBERTa-wwm).

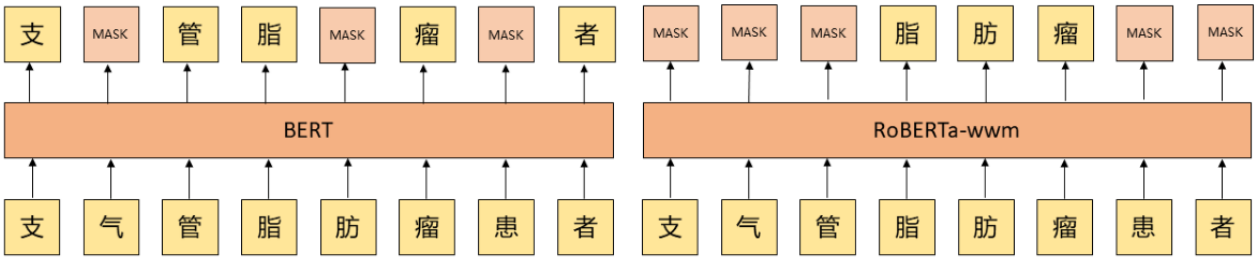
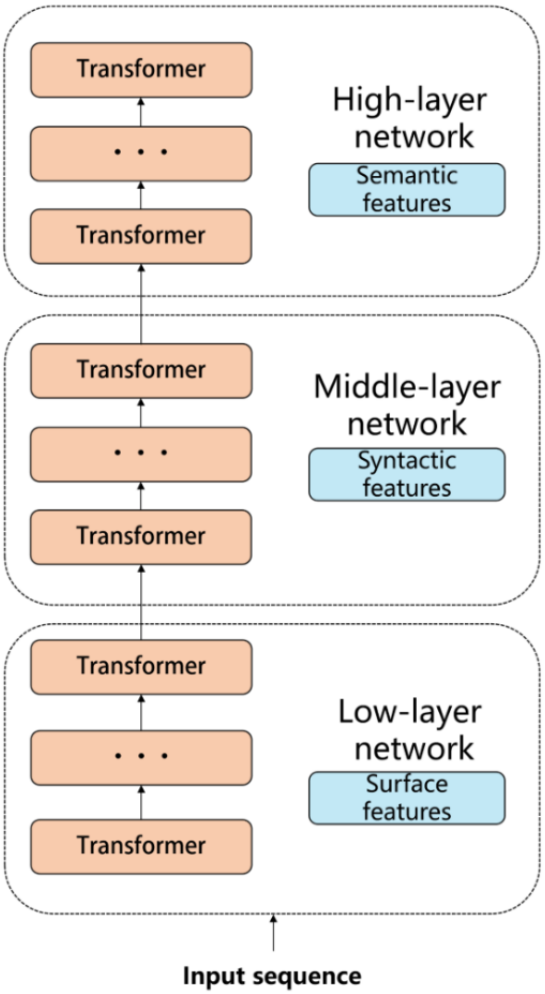


Figure 4. Coding representation of Transformer with 12 layers of Bidirectional Encoder Representation from Transformers model.



Assume that the text input sequence $seq = (x_1,x_2,x_3, \ldots, x_n)$, where n is the total length of the character contained in the sequence; x_i is the i^{th} character of the input sequence; and the fusion formula is as follows:

$$v_i^{RoBERTa\text{-}wwm} = Dense_{unit=512}(x_i)\{\sum_{i=1}^{n}\alpha_c \times h_c\}, (c \in [1,12]) \quad \textbf{(4)}$$

$v_i^{(RoBERTa\text{-}wwm)}$ is the output representation by the RoBERTa-wwm model with dynamic fusion for the current character x_i; h_c is the output representation by each transformer layer of the RoBERTa-wwm model, and α_c is the output weight value assigned to each layer by RoBERTa-wwm.

Fine-Grained Semantic Feature

5-Stroke Code

The 5-stroke code is a typical semantic code, which encodes Chinese characters according to strokes and structures. Currently, it is widely used to code Chinese characters. The expression of the 5-stroke code may inevitably repeat with the phonological code, for example, the 5-stroke code for "亦 (also)" is "you," while the phonological code for "亦 (also)" is also "you" [62]. Hence, we combined the 5-stroke code and Zheng code to compensate for the encoding deficiency. We used the 5-stroke code in Zdic.net to vectorize the Chinese characters using the following formulas:

$$p = f_{fc}\ (seq)\ \textbf{(5)}$$

$$vifc=efc\ (pi),\ (i \in Z \cap i \in\ [1,n])\ \textbf{(6)}$$

where f_{fc} represents the function that maps the input character sequence into the 5-stroke code and v_i^{fc} represents the 5-stroke code vector corresponding to x_i.

Zheng Code

The Zheng code was created by famous Chinese literature professors as per the strokes and roots of Chinese characters through in-depth research on the patterns and structures of Chinese characters. The early Microsoft operating system in Chinese adopted the Zheng code as the built-in code. This indicates that Zheng code is a scientific coding of Chinese characters. Chinese characters with similar codes may contain related semantic information. Hence, the potential semantic relationship of text may be found by mining the structural information of Chinese characters using Zheng code. The Zheng code was vectorized as the 5-stroke code and has the following formulas:

$$p = f_{zc}\ (seq)\ \textbf{(7)}$$

$$vizc=ezc\ (pi),\ (i \in Z \cap i \in\ [1,n])\ \textbf{(8)}$$

where f_{zc} represents the function that maps the input character sequence into the Zheng code and v_i^{zc} represents the Zheng code vector corresponding to x_i.

Phonological Code

Over 90% of Chinese characters are picto-phonetic characters [63]. Hence, pronunciation plays an important role in the semantic expressions of Chinese characters. We used the Pypinyin toolkit to vectorize the phonological code of Chinese characters using the following formulas:

$$p = f_{pc}\ (seq)\ \textbf{(9)}$$

$$vipc=epc\ (pi),\ (i \in Z \cap i \in\ [1,n])\ \textbf{(10)}$$

where f_{pc} represents the function that maps the input character sequence into the phonological code and v_i^{pc} represents the phonological vector corresponding to x_i.

Stroke Code

Chinese characters with similar strokes may have similar meanings. The strokes of each Chinese character were encoded in ZDIC [59], where 1, 2, 3, 4, and 5 represent the horizontal stroke, vertical stroke, left-falling stroke, right-falling stroke, and turning stroke, respectively. We transformed the stroke code into a 5-dimension vector, where each dimension was the corresponding number of strokes. The stroke code was vectorized in the same manner as the 5-stroke code and has the following formulas:

$$p = f_{sc}\ (seq)\ \textbf{(11)}$$

$$visc=esc\ (pi),\ (i \in Z \cap i \in\ [1,n])\ \textbf{(12)}$$

where f_{sc} represents the function that maps the input character sequence into the stroke code and v_i^{sc} represents the stroke code vector corresponding to x_i.

To extract the fine-grained semantic features of Chinese characters deeply, we trained the character features using CNNs. The character features were trained by 2 convolutions with a kernel of 3 and *ReLU* function as well as max pooling of 2×2, where the number of output channels was the dimension of each feature vector. Finally, the 32-dimension Chinese character vector was obtained through a full connection in the dense layer, as shown in Figure 2.

Image Feature

Chinese characters have been derived from pictographic symbols since ancient times, and characters with similar symbolic appearances have similar image features. However, the fonts of Chinese characters have changed a lot over time. Simplified characters have lost much pictographic information compared with complex characters. Therefore, Cui et al [64] used Chinese character images to extract Chinese character features and achieved better performance. Wu et al [65] tried different character fonts and found that the best result was obtained by using the *NotoSansCJKsc-Regular* font. On the basis of these findings, we used the Python Imaging Library to convert *NotoSansCJKsc-Regular* Chinese characters into black-and-white images and extracted image features by 2D CNN in depth as per the following formulas:

$$e^{if\_1}=(Max\ pooling\ 1\ (Conv\ 1\ (K \otimes H))\ \textbf{(13)}$$

$$e^{if\_2}=(Max\ pooling\ 2\ (Conv\ 2\ (K \otimes H)\ \textbf{(14)}$$

$$v_i^{if}=\text{Dense}\ (e^{if\_2})\ \textbf{(15)}$$

where K is a kernel; H is the original embedded image matrix; *Conv 1*, *Max pooling 1*, *Conv 2*, and *Max pooling 2* are the first convolutions with a kernel of 3 and channel of 8, the first max pooling with the kernel of 2×2, the second convolution with the kernel of 3 and channel of 32, and the second max pooling with the kernel of 2×2, respectively; e^{if_1} is the result after the first convolution; e^{if_2} is the result after the second convolution; Dense is the process of realizing the full connection; and v_i^{if} is the final 128-dimension Chinese character image vector trained by convolution, as shown in Figure 2.

Finally, the multisemantic features $v_i^{RoBERTa\text{-}wwm}$, v_i^{fc}, v_i^{zc}, v_i^{pc}, v_i^{sc}, and v_i^{if} were embedded by the array Concat function. The formula used is as follows:

$$v_i^{input}=\text{Concat}\ (v_i^{RoBERTa\text{-}wwm}, v_i^{fc}, v_i^{zc}, v_i^{pc}, v_i^{sc}, v_i^{if})\ \textbf{(16)}$$

BiLSTM Layer

The role of BiLSTM [66] is essential in NER. As shown in Figure 1, the forward LSTM and backward LSTM are responsible for memorizing the previous and subsequent text information, respectively. By combining the 2, contextual information can be obtained simultaneously, which helps to capture the bidirectional semantic dependency information in the text. The formulas used are as follows:

$$h_i^{forward}=LSTM^{forward}(\alpha^{<i-1>}, x_i) \textbf{ (17)}$$

$$h_i^{backward}=LSTM^{backward}(\alpha^{<i-1>}, x_i) \textbf{ (18)}$$

$$h_i=[h_i^{forward}; h_i^{backward}] \textbf{ (19)}$$

where $\alpha^{<i>}$ represents the hidden layer state of the current memory cell; $LSTM^{forward}$ is the feature representation from front to back; $LSTM^{backward}$ is the feature representation from back to front; $h_i^{forward}$ is the forward semantic information obtained through the forward LSTM at the *i*-th character position; $h_i^{backward}$ is the backward semantic information obtained through the backward LSTM at the *i*-th character position; and h_i represents a combination of hidden states in both.

CRF Layer

The BiLSTM can be used to handle contextual relationships. However, it cannot consider the dependencies between tags. Therefore, it is necessary to add a constraint relation for the final predicted label by using the CRF [67] layer to ensure the predicted label rationality. Given an input sequence where $X=\{x_1,x_2,...,x_n\}$, we assume that the training output label sequence is $Y=\{y_1,y_2,...,y_n\}$, where n is the number of model labels. The sequence score of the label and the probability of the label sequence y are calculated as follows:

$$P(y|X)=e^{\sum i=1}(Z_{yi,yi+1}+P_{i+1,yi+1}) / (\sum_{(y\tilde{}\in Yx)} e^{\sum i=1}(Z_{yi,yi+1}+P_{i+1,yi+1}) \textbf{ (20)}$$

where Z is the transfer matrix; $Z_{yi,yi+1}$ is the score of the label transfer from y_i to y_{i+1}; $P_{i+1,yi+1}$ is the score of label y_{i+1} corresponding to the $i+1$th character of the input sequence; Y_x is the set of all possible label sequences. The final label of the output sequence is the set of labels with the highest probability.

Finally, we predicted the best label sequences by using the Viterbi algorithm [68] with the following formula:

$$y^*=argmax(s(X, y)) \textbf{ (21)}$$

Results

To get convincing experimental results, we ran each model 5 times and calculated the average precision, average recall, and average F_1-score.

Performance Comparison With Ensemble Models

To verify the validity of the model, we compared our model with the existing ensemble models BiLSTM-CRF, ELMo-Lattice-LSTM-CRF, ELMo-BiLSTM-CRF, all CNNs, ELMo-encoder from transformer-CRF, and multigranularity semantic dictionary and multimodal tree-NER on Yidu-S4K and self-annotated data sets, and the results are shown in Table 4. The F_1-scores of the experimental model on the Yidu-S4K data set were 18.31%, 4.15%, 4.26%, 4.12%, 3.69%, and 2.59% higher than those of the BiLSTM-CRF, all CNNs, ELMo-Lattice-LSTM-CRF, ELMo-BiLSTM-CRF, ELMo-encoder from transformer-CRF, and multigranularity semantic dictionary and multimodal tree-NER models, respectively. On the self-annotated data set, it was 5.14% higher than that of the BiLSTM-CRF. The results showed that the performance of the experimental model is superior to that of the existing model.

Table 4. Performance comparison of ensemble models on the Yidu-S4K and self-annotated data sets.

Data set and model	Precision (%)	Recall (%)	F_1-score (%)
Yidu-S4K			
BiLSTM-CRF[a] [64]	69.43	72.58	70.97
ACNN[b] [69]	83.07	87.29	85.13
ELMo[c]-lattice-LSTM-CRF [70]	84.69	85.35	85.02
ELMo-BiLSTM-CRF [41]	—[d]	—	85.16
ELMo-ET[e]-CRF [71]	82.08	86.12	85.59
MSD_DT_NER[f] [72]	86.09	87.29	86.69
Our model	90.37	88.22	89.28
Self-annotated			
BiLSTM-CRF	81.98	77.10	79.47
Our model	84.24	84.99	84.61

[a]BiLSTM-CRF: Bidirectional Long Short-Term Memory-conditional random fields.

[b]ACNN: all convolutional neural network.

[c]ELMo: Embeddings from Language Models.

[d]Not available.

[e]ET: encoder from transformer.

[f]MSD_DT_NER: multigranularity semantic dictionary and multimodal named entity recognition.

Performance Comparison With PLMs Related to BERT

The performance of the PLM, BERT, is a milestone in natural language processing. To verify the BERT robust version's validity of the RoBERTa-wwm model, we compared our model with the existing ensemble models with the BiLSTM-CRF, BERT-BiLSTM-CRF, and RoBERTa-wwm-BiLSTM-CRF on Yidu-S4K and self-annotated data sets, and the results are shown in Table 5. The F_1-scores of the experimental model on the Yidu-S4K data set were 18.31%, 2.99%, and 0.82% higher than those of the BiLSTM-CRF, BERT-BiLSTM-CRF, and RoBERTa-wwm-BiLSTM-CRF models, respectively, and 5.14%, 2.95%, and 1.07% higher on the self-annotated data set, respectively.

Table 5. Performance comparison of PLMs[a] on the Yidu-S4K and self-annotated data sets.

Data set and model	Precision (%)	Recall (%)	F_1-score (%)
Yidu-S4K			
BiLSTM[b]-CRF[c] [64]	69.43	72.58	70.97
BERT[d]-BiLSTM-CRF	89.07	83.67	86.29
RoBERTa-wwm[e]-BiLSTM-CRF	90.08	86.90	88.46
Our model	90.37	88.22	89.28
Self-annotated			
BiLSTM-CRF	81.98	77.10	79.47
BERT-BiLSTM-CRF	82.48	80.86	81.66
RoBERTa-wwm-BiLSTM-CRF	84.23	82.86	83.54
Our model	84.24	84.99	84.61

[a]PLM: pretrained language model.

[b]BiLSTM: Bidirectional Long Short-Term Memory.

[c]CRF: conditional random fields.

[d]BERT: Bidirectional Encoder Representation from Transformers.

[e]RoBERTa-wwm: Robustly Optimized Bidirectional Encoder Representation from Transformers Pretraining Approach Whole Word Masking.

Performance Comparison of Each Entity

To comprehensively evaluate our model, we calculated the F_1-score for each entity type on the Yidu-S4K and self-annotated data sets, as shown in Tables 6 and 7. The F_1-score of our model on the Yidu-S4K data set for each of the 6 entity categories, except for the Image entity, increased by 0.2% to 7.6% compared with the data listed in the tables. The F_1-score for the Image entity was 0.35% lower than that of the ELMo-BiLSTM-CRF model. However, the F_1-scores for the Laboratory entity and Operation entity were 7.6% and 7.54% higher than those of the ELMo-BiLSTM-CRF model, respectively. The overall F_1-score was 4.12% higher than that of the ELMo-BiLSTM-CRF model. For the self-annotated data set, our model improved each entity in 7 categories ranging from 0.09% to 14.49% over the listed data, with a greater improvement for Instrument entities.

Table 6. Performance comparison of each entity category on the Yidu-S4K data set.

Model	F_1-score for each category (%)						
	All	Disease	Anatomy	Image	Laboratory	Medicine	Operation
ELMo[a]-BiLSTM[b]-CRF[c] [41]	85.16	82.81	85.99	88.01	75.65	94.49	86.79
BERT[d]-BiLSTM-CRF	86.29	87.14	86.36	83.43	77.98	89.46	93.11
BERT-wwm[e]-BiLSTM-CRF	87.12	86.18	85.47	81.52	79.69	90.14	92.49
RoBERTa[f]-wwm-BiLSTM-CRF	88.46	87.71	87.01	86.69	82.36	93.22	92.87
Our model	89.28	87.91	87.47	87.66	83.25	94.98	94.33

[a]ELMo: Embeddings from Language Models.

[b]BiLSTM: Bidirectional Long Short-Term Memory.

[c]CRF: conditional random fields.

[d]BERT: Bidirectional Encoder Representation from Transformers.

[e]wwm: Whole Word Masking.

[f]RoBERTa: Robustly Optimized Bidirectional Encoder Representation from Transformers Pretraining Approach.

Table 7. Performance comparison of each entity category on the self-annotated data set.

Model	F_1-score for each category (%)							
	All	Disease	Symptoms	Anatomy	Examination	Instrument	Medicine	Operation
BERT[a]-BiLSTM[b]-CRF[c]	81.66	81.33	85.87	83.86	90.36	60.38	89.72	79.75
BERT-wwm[d]-BiLSTM-CRF	81.58	74.91	83.89	81.23	88.84	54.76	85.63	68.49
RoBERTa[e]-wwm-BiLSTM-CRF	83.54	81.99	86.69	84.68	91.21	66.01	91.04	81.17
Our model	84.61	82.34	86.93	85.62	91.30	69.25	91.28	82.49

[a]BERT: Bidirectional Encoder Representation from Transformers.

[b]BiLSTM: Bidirectional Long Short-Term Memory.

[c]CRF: conditional random fields.

[d]wwm: Whole Word Masking.

[e]RoBERTa: Robustly Optimized Bidirectional Encoder Representation from Transformers Pretraining Approach.

Ablation Analysis

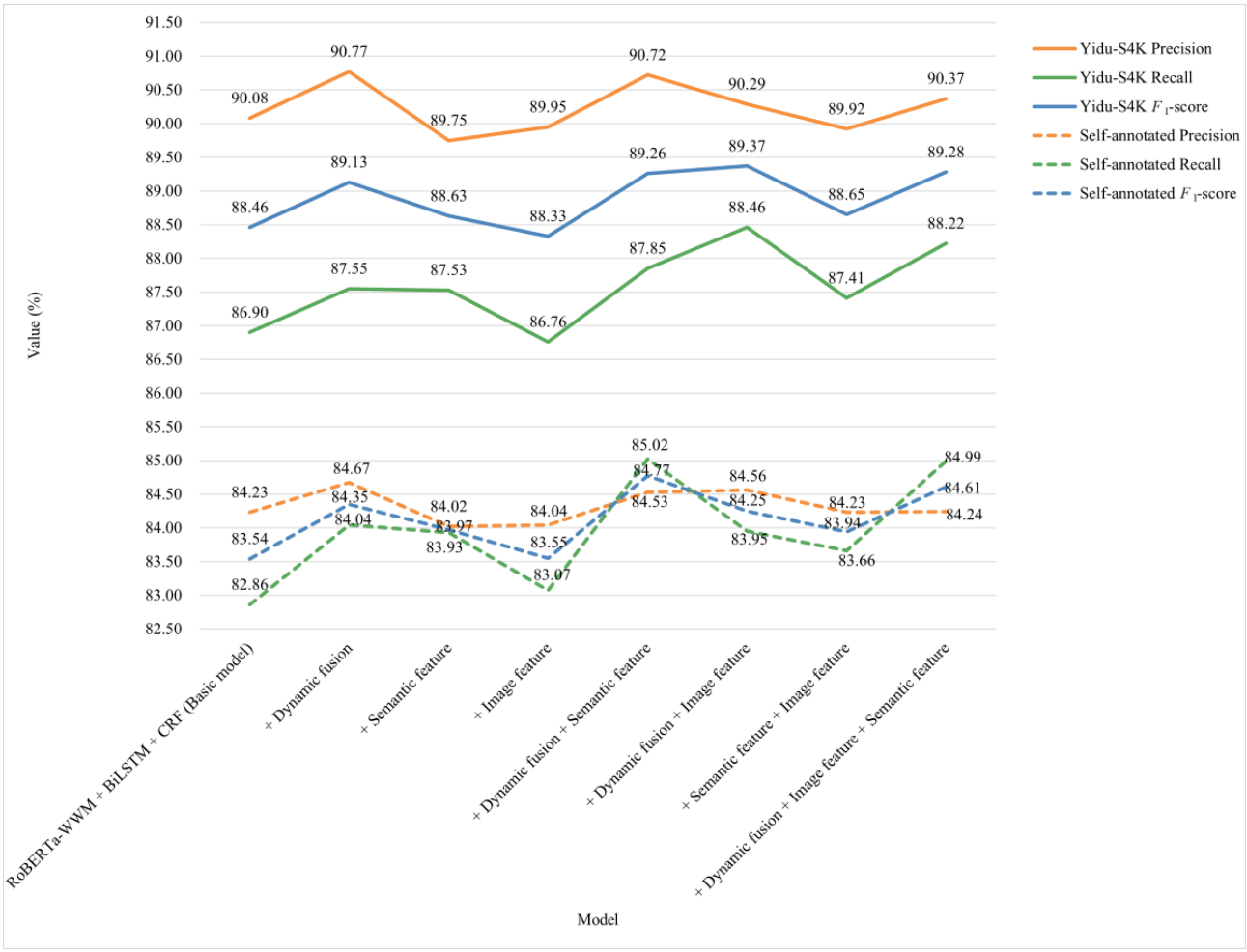
Ablation Experiments for Multisemantic Features

To verify the fine-grained semantic features and image features of Chinese characters, dynamic fusion was effective. We used the RoBERTa-wwm-BiLSTM-CRF model as the baseline to perform ablation experiments for the above contents on 2 EMR data sets, and the results are shown in Figure 5.

The performance of the model was significantly improved with the dynamic fusion of RoBERTa-wwm. After incorporating the semantic features of Chinese characters into the model alone, the overall performance of the model was not as high as that after dynamic fusion. However, the performance on both data sets was superior to that of the baseline. The performance of the model was unstable when image features of Chinese characters were added to the model alone. On the Yidu-S4K data set, the model's performance was inferior to that of the baseline, whereas on the self-annotated data set, the model's performance only improved slightly. After adding the semantic and image features of Chinese characters to the model, the performance of the model on the Yidu-S4K data set was superior to that of the baseline. Furthermore, it was better than that of the model with semantic or image features of Chinese characters alone. The performance of the model on the self-annotated data set was superior to that of the baseline and better than that of the model with the image features of Chinese characters alone. When the model combined dynamic fusion with the semantic

features and image features of Chinese characters, it was found that the performance of the model was significantly improved on the 2 data sets. Dynamic fusion with image features of Chinese characters showed the best comprehensive performance on the Yidu-S4K data set, whereas dynamic fusion with semantic features of Chinese characters achieved the best comprehensive performance on the self-annotated data set. After combining the semantic and image features of the Chinese characters and dynamic fusion, it was noted that the performance of the model was superior to that of the baseline. Because the quality of the self-annotated EMRs is inferior to that of the public Chinese EMRs corpus and the self-annotated data set contains a wider coverage of departments, the comprehensive effect of the self-annotated data set is lower than that of the YiduS4K data set in Figure 5.

Figure 5. The results of ablation experiments for mutisemantic features on the Yidu-S4K and self-annotated data sets. BiLSTM: Bidirectional Long Short-Term Memory; CRF: Conditional Random Fields; RoBERTa-wwm: Robustly Optimized Bidirectional Encoder Representation from Transformers Pretraining Approach Whole Word Masking.



Ablation Experiments for Fine-Grained Semantic Features

The fine-grained semantic features of Chinese characters used in this study included the 5-stroke code, Zheng code, phonological code, and stroke code. To verify the effectiveness of these features, we used the RoBERTa-wwm-BiLSTM-CRF model as the baseline to perform ablation experiments for the 4 features on the 2 EMR data sets, and the results are shown in Figure 6 and Figure 7.

Figure 6. The results of ablation experiments for fine-grained semantic features on the Yidu-S4K data set. BiLSTM: Bidirectional Long Short-Term Memory; CRF: conditional random fields; RoBERTa-wwm: Robustly Optimized Bidirectional Encoder Representation from Transformers Pretraining Approach Whole Word Masking.

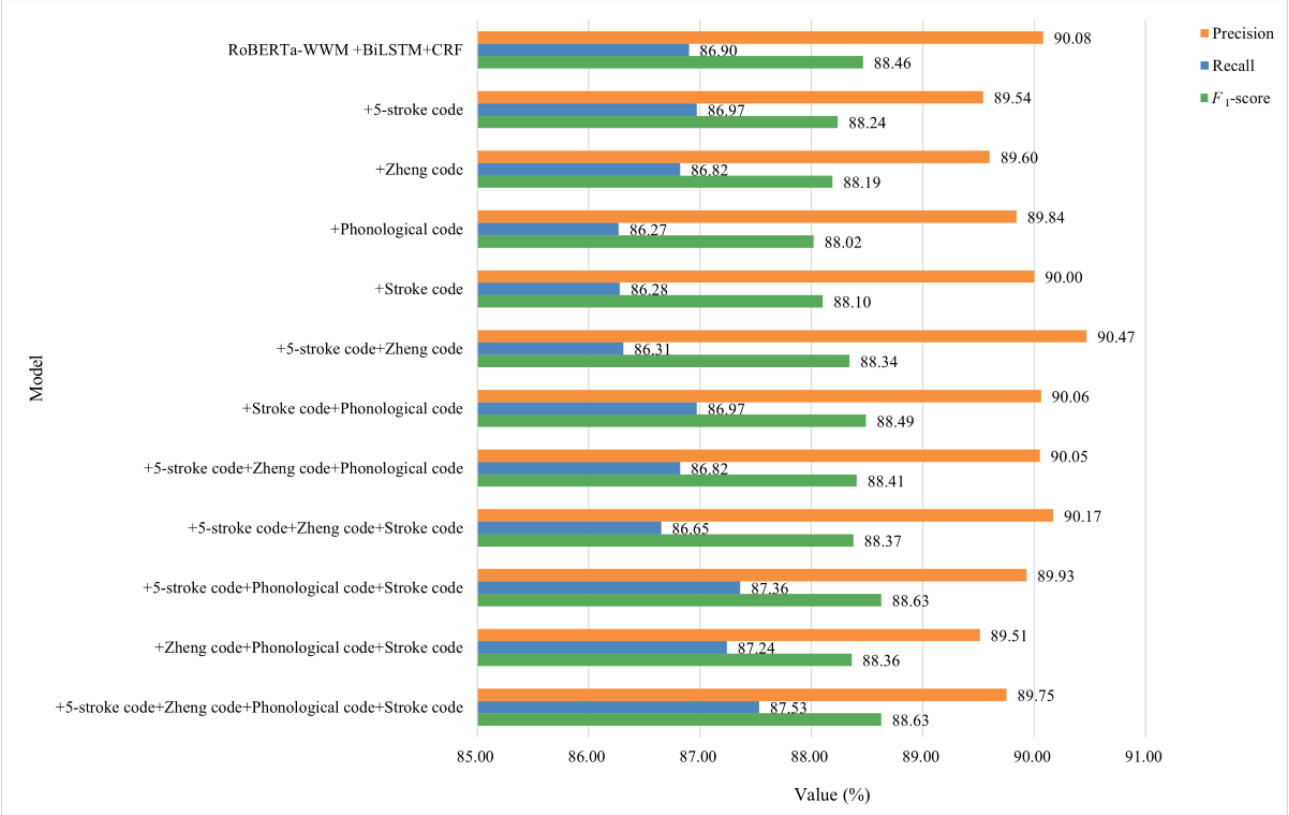
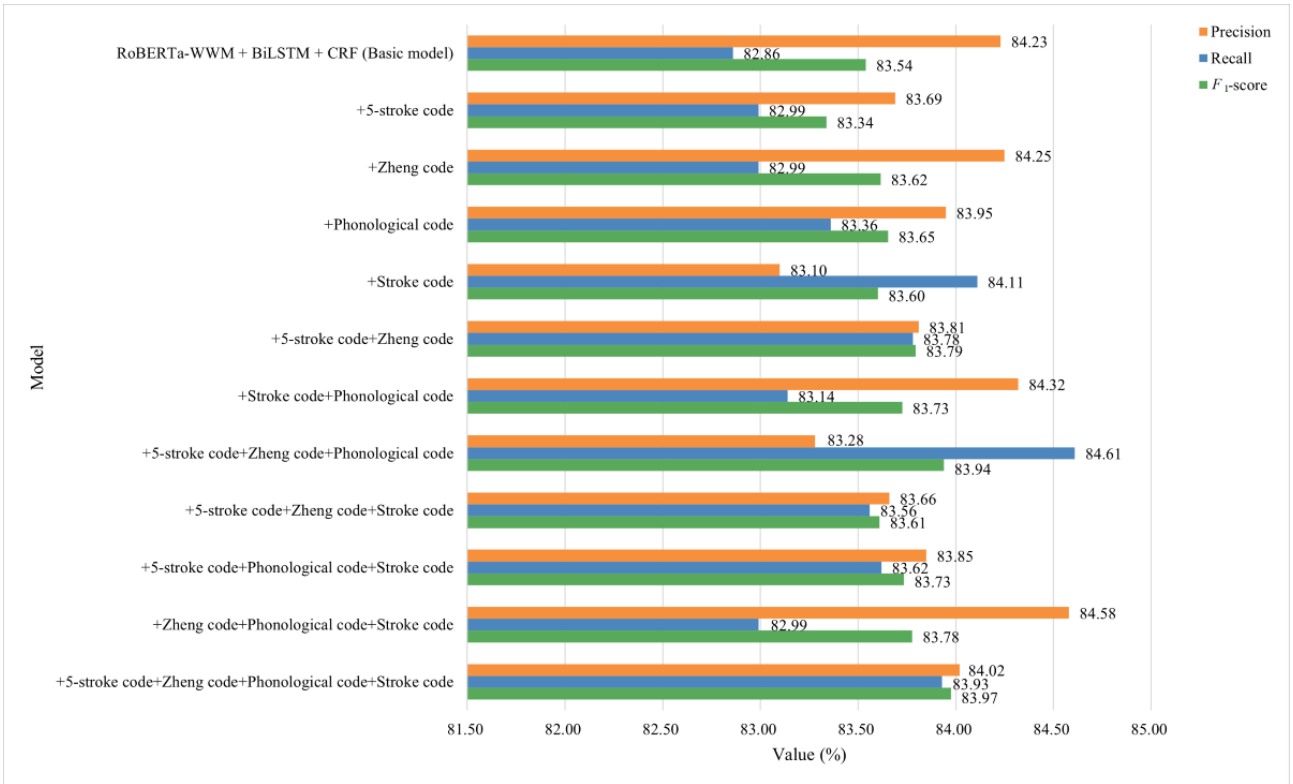


Figure 7. The results of ablation experiments for fine-grained semantic features on the self-annotated data set. BiLSTM: Bidirectional Long Short-Term Memory; CRF: conditional random fields; RoBERTa-wwm: Robustly Optimized Bidirectional Encoder Representation from Transformers Pretraining Approach Whole Word Masking.



The F_1-score of the model on the Yidu-S4K data set ranked in the top 2 for the 5-stroke code and Zheng code, whereas the F_1-score on the self-annotated data set ranked in the top 2 for the phonological code or Zheng code. The performance of the

model combining 2 features (the combination of 5-stroke code and Zheng code or the combination of phonological code and stroke code) was better than that of the model with only 1 feature, regardless of the data set. On the Yidu-S4K data set, the model combining 5-stroke code+phonological code+stroke code showed the best comprehensive performance, followed by the combinations of 5-stroke code+Zheng code+phonological code, 5-stroke code+Zheng code+stroke code, and Zheng code+phonological code+stroke code. On the self-annotated data set, the model combining the 5-stroke code+Zheng code+phonological code showed the best comprehensive performance, followed by Zheng code+phonological code+stroke code, 5-stroke code+phonological code+stroke code, and 5-stroke code+Zheng code+stroke code. On the Yidu-S4K data set, only the model combining the 5-stroke code+phonological code+stroke code showed a comprehensive performance superior to that of the baseline. However, on the self-annotated data set, the comprehensive performance of all combinations was superior to that of the baseline. The performance of the model combining 3 features was less stable. The model combined 4 features on the Yidu-S4K and self-annotated data sets and achieved the best comprehensive performance among all the combinations.

Error Analysis

From Tables 6 to Table 7, our model improved the entity recognition performance of each entity category to different degrees. However, the entity recognition effect differs for each category. The F_1-scores of Disease, Anatomy, Image, Laboratory, Medicine, and Operation entity recognition on the Yidu-S4K data set were 87.91%, 87.47%, 87.66%, 83.25%, 94.98%, and 94.33%, respectively. The F_1-scores of Disease, Symptoms, Anatomy, Examination, Instrument, Medicine, and Operation entity recognition on the self-annotated data set were 82.34%, 86.93%, 85.62%, 91.31%, 69.25%, 91.28%, and 82.49%, respectively. On the Yidu-S4K data set, the precision of Laboratory entity recognition was the lowest, followed by the Anatomy entity, Image entity, and Disease entity. On the self-annotated data set, the precision of Instrument entity recognition was the lowest, followed by the Disease entity, Anatomy entity, and Operation entity. We concluded the following 7 main causes of the errors that occurred based on a review of the data set and model prediction results, as shown in Table 8.

We strictly controlled the annotation quality of both data sets. Hence, the probability of causes (1-3) was relatively low. Causes (4-6) were more likely to occur, and cause (7) mainly occurred on some entities that were less common or had fewer training samples.

Table 8. Different types of errors on 2 data sets.

Types of errors Illustrations	Example
(1) Annotation error	
1. Some manually annotated entities contained punctuation marks unrelated to the entities.	For instance, some Laboratory entities, like "PLT[a]," "NEUT[b]," and "CAE[c]," on the Yidu-S4K data set contained commas, which were correctly recognized as "PLT," "NEUT," and "CAE" in our model.
2. A few entity categories were confused.	For example, "PET-CT[d]" was manually annotated as a Laboratory entity on the Yidu-S4K data set, but our model correctly predicted as an Image entity.
(2) Inconsistent annotation	
The inconsistent annotation will affect the accuracy of machine learning.	On the Yidu-S4K data set, the character "下 (below)" expressing orientation of "剑突下 (below xiphoid)" was not annotated, and the character "部 (part)" expressing the part of "咽喉部 (hypopharynx)" was also not annotated. Most of the characters expressing specific locations were annotated.
(3) Missing annotation	
The missing annotated entity will also affect the overall effect of the model.	The Disease entity "窦性心律 (sinus rhythm)" was missed annotated on the Yidu-S4K data set, and the Medicine entity "生理盐水 (normal saline)" was missed annotated on the self-annotated data set.
(4) Entity with a non-Chinese character symbol	
Figures, letters, and other symbols cannot be extracted with more semantic features than Chinese characters. Hence, it may be difficult to recognize entities with symbols other than Chinese characters in the Chinese corpus.	The failure to recognize the non-Chinese character entities, like the Laboratory entity "AFP[e]" on the Yidu-S4K data set and the Examination entity "HCG[f]," on the self-annotated data set, so did such situations as the Medicine entity "VP[g]-16" was recognized as "VP-," and "50%葡萄糖 (50% glucose)" as "葡萄糖 (glucose)" on the Yidu-S4K data set.
(5) Presence of nested entities	
On the Yidu-S4K data set, the Disease entity and Image entity might contain the Anatomy entity.	For example, the Disease entity "二尖瓣后叶钙化 (posterior mitral valve leaflet calcification)" was recognized as the Anatomy entity "二尖瓣 (bicuspid)," and the Image entity "腹部彩超 (abdominal color doppler ultrasound)" was recognized as the Anatomy entity "腹部 (abdominal)."
On the self-annotated data set, entity nesting is more severe, the Disease entity, Examination entity, and Instrument entity might contain the Anatomy entity, and the Instrument entity might contain the Operation entity.	For example, the Disease entity "内踝骨折 (ankle fracture)" was recognized as the Anatomy entity "内踝 (medial malleolus)," the Examination entity "骨髓组织病理 (bone marrow histopathology)" was recognized as the Anatomy entity "骨髓 (bone marrow)," the Instrument entity "胸部支具 (chest brace)" was recognized as the Anatomy entity "胸(chest)," and the Instrument entity "左胸引流管 (left thoracic drainage tube)" was recognized as the Operation entity "左胸引流 (left thoracic drainage)."
(6) More entities with mixed representation	
Entity composition is more complex, mixed representations occur more often.	The Medicine entity "奥沙利铂 (乐沙定) (Oxaliplatin [Eloxatin])" on the Yidu-S4K data set was recognized as "奥沙利铂 (Oxaliplatin)" and "乐沙定 (Eloxatin)," respectively, the Disease entity "CD5[h]+弥漫大B细胞淋巴瘤 (白血病期)" on the self-annotated data set was recognized as "CD" and "弥漫大B细胞淋巴瘤 (白血病期) (diffuse large B-cell lymphoma [Leukemia stage])," and the Examination entity "肥达、外斐反应 (Widal, well-felix reaction)" on the self-annotated data set was recognized as "肥达 (Widal)" and "外斐反应 (well-felix reaction)," respectively.
(7) Insufficient entity training data	
In the case of insufficient training samples, the machine may provide inadequate training for entities, so that the machine cannot fully learn the features of such entities, failing to recognize many entities.	On the self-annotated data set, the number of Instrument entities was less than that of other categories (Table 2), accounting for only 1.52% of the total, those entities might never appear in the training data set, such as "针筒 (syringe)," "微导管 (microtubule)," "550px碳钢钻头 (550px carbon steel drill bit)," etc.

[a]PLT: platelet count.

[b]NEUT: neutrophil count.

[c]CAE: carcinoembryonic antigen.

[d]PET-CT: positron emission tomography-computed tomography.

[e]AFP: alpha fetoprotein.

[f]HCG: human chorionic gonadotropin.

[g]VP: etoposide.

[h]CD5: a differentiation antigen, cluster of differentiation 5.

Discussion

Principal Findings

In this study, we developed a Chinese CNER method based on multisemantic features. The method extracted the semantic features of text using the RoBERTa-wwm model after dynamic fusion, extracted the fine-grained semantic features of Chinese characters by 1D CNN, and converted Chinese characters into square images to extract the image features of the simplified Chinese characters from another modality by 2D CNN. We conducted a series of experiments to evaluate the model's performance on the Yidu-S4K data set and self-annotated data set; the results showed that the F_1-scores of the proposed model in this study were 89.28% and 84.61% on the 2 data sets, respectively. The model showed a higher and more stable performance in all experiments and could help recognize entities in most categories. Furthermore, its migrative property and adaptability to different data were acceptable. We also demonstrated that multisemantic features were effective through 2 ablation experiences and analyzed the error cases of NER, which might provide a basis for subsequent studies and standardization of the corpus.

Compared with ensemble models, for the BiLSTM-CRF model, the representation information of characters was obtained with the help of a vector look-up table. However, the information obtained by this method was too simple to excavate the text's semantic meaning or solve problems such as the polysemy of words. Hence, the model did not perform well. Kong et al [69] constructed a multilayer CNN to obtain short-term and long-term contextual information, and the attention mechanism was used to calculate the weight distribution in each hidden layer so that the features of each coding layer could be fully extracted and used to improve the entity recognition performance. However, this model required numerous radical and dictionary features to complete the semantic supplement of the context. Li et al [70] proposed an ELMo-Lattice-LSTM-CRF model. The ELMo word dynamic representation model could learn complicated word features and the context-based changes of these features, while the lattice structure provided extra entity boundaries and other semantic information for CNER of EMRs through the Word2Vec model and dictionaries. Li et al [41] proposed an ELMo-BiLSTM-CRF model that improved the semantic recognition ability of the machine for text. It reduced problems, such as word polysemy, when compared with the BiLSTM-CRF model and reduced the computational complexity of the lattice structure compared with the ELMo-Lattice-LSTM-CRF model. Moreover, this model could fully use contextual information by replacing LSTM with BiLSTM. Wan et al [71] fine-tuned the ELMo model based on EMRs to achieve embedding for domain-specific text and then used a transformer as an encoder to alleviate the long context–dependent problems and finally achieved CNER through CRF decoding. Wang et al [72] proposed a model for NER based on the LSTM-CRF model by storing and merging characters, words, and other features. However, as the text embedding process of this method is more complicated, it is necessary to create dictionaries of characters and words to obtain multigranularity text features at first and then store and merge the obtained features using a tree structure to achieve text embedding. These methods have achieved a few good results, but our proposed method is still competitive and has the best performance among all the models, as shown in Table 4.

Compared with PLMs related to BERT, both the BERT-BiLSTM-CRF and BiLSTM-CRF models could obtain word-level vector representations. However, the word-level vector obtained by BERT contained rich contextual characteristics, including morphology, syntax, semantics, location, and other important semantic information, which can directly improve the task performance by replacing the lattice structure and complicated text representation methods in Table 4, such as dictionaries of characters and words. Compared with BERT, RoBERTa-wwm used more data for pretraining, and the dynamic wwm allows itself to flexibly learn word-level representation information, which compensates for the shortcomings that BERT can only obtain character-level representation. Thus, richer word-based text representation information could be obtained. Combined with the experimental results in Table 4, the RoBERTa-wwm-BiLSTM-CRF model, without introducing features, outperformed the other ensemble models. Therefore, using the PLM RoBERTa-wwm with a whole word mask can effectively improve the Chinese CNER performance, thus avoiding the use of complex text embedding and feature embedding methods.

In addition, 2 ablation experiments showed that different features and means lead to different degrees of improvement in the semantic comprehension ability of the model. Multisemantic features could help the machine to obtain richer semantic information, whereas dynamic fusion could fully recognize and used the representation information so that the model performance could be comprehensively improved. Considering the heterogeneity among data, using 1 method alone or both methods may affect the generalization ability of the model. In this study, the model combining the fine-grained semantic features and image features of Chinese characters and dynamic fusion might not show the best performance. However, it was more universal and could maintain the performance at a relatively high level compared with other experimental models. Furthermore, introducing more feature engineering was conducive to fully mining the semantic information of text connotation with the help of fine-grained semantic information contained in Chinese characters and improving the performance of the model on different data sets through the cross-complementarity of different features in a relatively stable manner.

To reduce the error rate of entity recognition, specifically for human-caused errors, we could try to avoid these problems by further improving the annotation quality. For the data special characteristics or data defects, the errors might be reduced by medical knowledge, medical dictionaries, and some rules, regardless of the lack of training data.

Limitations and Future Work

The limitation of this study was that the ratio of the 6 entity types on the Yidu-S4K data set did not exactly follow 7:3, such that the ratio of the training set to test set for disease entities is approximately 0.7610:0.2390; the ratio of the training set to test

set for medicine entities is approximately 0.7898:0.2102; and the ratio of the training set to test set for all entities is approximately 0.7463:0.2537. The unbalanced data of different entity types in the training and test sets caused a performance bias. Although the ratio of the training set to the test set of the EMRs was 7:3, we could not ensure that the number of entities of each type in each EMR in the training set and test set remained at a similar ratio.

In the future, we may focus on the recognition of a specific entity type in EMRs to improve the CNER performance. In addition, we will incorporate other prior medical knowledge or assign different weights to the Chinese character semantic features and image features, such as using the attention mechanism to capture important features, to improve the performance of the model.

Conclusions

This study proposes a Chinese CNER method to learn a semantics-enriched representation of Chinese character features in EMRs to enhance the specificity and diversity of feature representations. The results showed that the model had state-of-the-art performance on 2 Chinese CNER data sets compared with existing models. We demonstrated that multisemantic features could provide richer and more fine-grained semantic information for Chinese CNER through the cross-complementarity of different semantic features. This enabled the model to learn a better feature representation and improve its generalization ability.

Acknowledgments

The authors would like to thank the YiduCloud for providing the Yidu-S4K corpora. This work was supported by the Science and Technology Innovation 2030 “New Generation Artificial Intelligence” major project “Research on the construction and application of knowledge system for Medical Artificial intelligence Services,” China (grant 2020AAA0104901).

Conflicts of Interest

None declared.

References

1. Shen L, Li Q, Wang W, Zhu L, Zhao Q, Nie Y, et al. Treatment patterns and direct medical costs of metastatic colorectal cancer patients: a retrospective study of electronic medical records from urban China. J Med Econ 2020 May;23(5):456-463. [doi: 10.1080/13696998.2020.1717500] [Medline: 31950863]
2. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform 2008:128-144. [Medline: 18660887]
3. Sang S, Yang Z, Liu X, Wang L, Lin H, Wang J, et al. GrEDeL: a knowledge graph embedding based method for drug discovery from biomedical literatures. IEEE Access 2018 Dec 12;7:8404-8415 [FREE Full text] [doi: 10.1109/access.2018.2886311]
4. Dogan RI, Lu Z. An improved corpus of disease mentions in PubMed citations. In: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing. 2012 Presented at: BioNLP '12; June 8, 2012; Montreal, Canada p. 91-99 URL: https://dl.acm.org/doi/10.5555/2391123.2391135
5. Saha S, Ekbal A, Sikdar UK. Named entity recognition and classification in biomedical text using classifier ensemble. Int J Data Min Bioinform 2015;11(4):365-391. [doi: 10.1504/ijdmb.2015.067954] [Medline: 26336665]
6. Shao Y, Hardmeier C, Tiedemann J, Nivre J. Character-based joint segmentation and POS tagging for Chinese using bidirectional RNN-CRF. In: Proceedings of the 8th International Joint Conference on Natural Language Processing. 2017 Presented at: IJCNLP' 17; November 27–December 1, 2017; Taipei, Taiwan p. 173-183 URL: https://aclanthology.org/I17-1018.pdf
7. Peng H, Cambria E, Zou X. Radical-based hierarchical embeddings for Chinese sentiment analysis at sentence level. In: Proceedings of the 30th International Florida Artificial Intelligence Research Society Conference. 2017 Presented at: FLAIRS '17; May 22–24, 2017; Marco Island, FL, USA p. 347-352 URL: https://sentic.net/radical-embeddings-for-chinese-sentiment-analysis.pdf
8. Shi X, Zhai J, Yang X, Xie Z, Liu C. Radical embedding: delving deeper to Chinese radicals. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. 2015 Presented at: IJCNLP '15; July 26-31, 2015; Beijing, China p. 594-598 URL: https://aclanthology.org/P15-2098.pdf [doi: 10.3115/v1/p15-2098]
9. Zhang WY. Contrastive Studies on English and Chinese vocabulary. Shanghai, China: Shanghai Foreign Language Education Press; 2010.
10. Liu JN. A study on the structure of Chinese characters. Jilin University. 2011. URL: https://kns.cnki.net/kcms2/article/abstract?v=3uoqIhG8C447WN1SO36whHG-SvTYjkCc7dJWN_daf9c2-IbmsiYfKhSrIbg5xu9MmzfA5stc78MC2nDoIg7_tb-KveYJIGpg&uniplatform=NZKPT [accessed 2023-04-17]
11. Tsuruoka Y, Tsujii J. Improving the performance of dictionary-based approaches in protein name recognition. J Biomed Inform 2004 Dec;37(6):461-470 [FREE Full text] [doi: 10.1016/j.jbi.2004.08.003]

12. Song M, Yu H, Han WS. Developing a hybrid dictionary-based bio-entity recognition technique. BMC Med Inform Decis Mak 2015;15(Suppl 1):S9 [FREE Full text] [doi: 10.1186/1472-6947-15-S1-S9] [Medline: 26043907]
13. Kraus S, Blake C, West SL. Information extraction from medical notes. Medinfo. 2007. URL: https://ils.unc.edu/phr/files/KrausBlakeWestMEDINFO2007.pdf [accessed 2022-05-15]
14. Tsai RT, Sung CL, Dai HJ, Hung HC, Sung TY, Hsu WL. NERBio: using selected word conjunctions, term normalization, and global patterns to improve biomedical named entity recognition. BMC Bioinformatics 2006 Dec 18;7 Suppl 5(Suppl 5):S11 [FREE Full text] [doi: 10.1186/1471-2105-7-S5-S11] [Medline: 17254295]
15. Khalifa M, Shaalan K. Character convolutions for arabic named entity recognition with long short-term memory networks. Comput Speech Lang 2019 Nov;58(C):335-346 [FREE Full text] [doi: 10.1016/j.csl.2019.05.003]
16. Na SH, Kim H, Min J, Kim K. Improving LSTM CRFs using character-based compositions for Korean named entity recognition. Comput Speech Lang 2019 Mar;54:106-121 [FREE Full text] [doi: 10.1016/j.csl.2018.09.005]
17. Yu H, An J, Yoon J, Kim H, Ko Y. Simple methods to overcome the limitations of general word representations in natural language processing tasks. Comput Speech Lang 2020 Jan;59:91-113 [FREE Full text] [doi: 10.1016/j.csl.2019.04.009]
18. Saha SK, Sarkar S, Mitra P. Feature selection techniques for maximum entropy based biomedical named entity recognition. J Biomed Inform 2009 Oct;42(5):905-911 [FREE Full text] [doi: 10.1016/j.jbi.2008.12.012] [Medline: 19535010]
19. Lee KJ, Hwang YS, Kim S, Rim HC. Biomedical named entity recognition using two-phase model based on SVMs. J Biomed Inform 2004 Dec;37(6):436-447 [FREE Full text] [doi: 10.1016/j.jbi.2004.08.012] [Medline: 15542017]
20. Ju Z, Wang J, Zhu F. Named entity recognition from biomedical text using SVM. In: Proceedings of the 5th International Conference on Bioinformatics and Biomedical Engineering. 2011 Presented at: ICBBE '11; May 10-12, 2011; Wuhan, China p. 1-4 URL: https://ieeexplore.ieee.org/document/5779984 [doi: 10.1109/icbbe.2011.5779984]
21. Shen D, Zhang JJ, Zhou G, Su J, Tan CL. Effective adaptation of hidden Markov model-based named entity recognizer for biomedical domain. In: Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine. 2003 Presented at: BioMed '03; July 11, 2003; Sapporo, Japan p. 49-56. [doi: 10.3115/1118958.1118965]
22. Zhou G, Su J. Named entity recognition using an HMM-based chunk tagger. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. 2002 Presented at: ACL '02; July 7-12, 2002; Philadelphia, PA, USA p. 473-480 URL: https://dl.acm.org/doi/abs/10.3115/1073083.1073163 [doi: 10.3115/1073083.1073163]
23. Liu J, Huang M, Zhu X. Recognizing biomedical named entities using skip-chain conditional random fields. In: Proceedings of the 2010 Workshop on Biomedical Natural Language Processing. 2010 Presented at: BioNLP '10; July 15, 2010; Uppsala, Sweden p. 10-18 URL: https://dl.acm.org/doi/pdf/10.5555/1869961.1869963
24. Skeppstedt M, Kvist M, Nilsson GH, Dalianis H. Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: an annotation and machine learning study. J Biomed Inform 2014 Jun;49:148-158 [FREE Full text] [doi: 10.1016/j.jbi.2014.01.012] [Medline: 24508177]
25. Gorinski PJ, Wu H, Grover C, Tobin R, Talbot C, Whalley H, et al. Named entity recognition for electronic health records: a comparison of rule-based and machine learning approaches. arXiv. Preprint posted online on June 5, 2019 2023. [doi: 10.48550/arXiv.1903.03985]
26. Suárez-Paniagua V, Rivera Zavala RM, Segura-Bedmar I, Martínez P. A two-stage deep learning approach for extracting entities and relationships from medical texts. J Biomed Inform 2019 Nov;99:103285 [FREE Full text] [doi: 10.1016/j.jbi.2019.103285] [Medline: 31546016]
27. Cocos A, Fiks AG, Masino AJ. Deep learning for pharmacovigilance: recurrent neural network architectures for labeling adverse drug reactions in Twitter posts. J Am Med Inform Assoc 2017 Jul 01;24(4):813-821 [FREE Full text] [doi: 10.1093/jamia/ocw180] [Medline: 28339747]
28. Wang Q, Zhou Y, Ruan T, Gao D, Xia Y, He P. Incorporating dictionaries into deep neural networks for the Chinese clinical named entity recognition. J Biomed Inform 2019 Apr;92:103133 [FREE Full text] [doi: 10.1016/j.jbi.2019.103133] [Medline: 30818005]
29. Bengio Y, LeCun Y. Convolutional networks for images, speech, and time series. In: Arbib MA, editor. The Handbook of Brain Theory and Neural Networks. Cambridge, MA, USA: MIT Press; Oct 1998:255-258.
30. Schmidhuber J. Deep learning in neural networks: an overview. Neural Netw 2015 Jan;61:85-117. [doi: 10.1016/j.neunet.2014.09.003] [Medline: 25462637]
31. Sundermeyer M, Schlüter R, Ney H. LSTM neural networks for language modeling. In: Proceedings of the 13th Annual Conference of the International Speech Communication Association. 2012 Presented at: INTERSPEECH '12; September 9-13, 2012; Portland, OR, USA p. 194-197 URL: https://www.isca-speech.org/archive/pdfs/interspeech_2012/sundermeyer12_interspeech.pdf [doi: 10.21437/interspeech.2012-65]
32. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. In: Proceedings of the Workshop at 1st International Conference on Learning Representations. 2013 Presented at: ICLR '13; May 2-4, 2013; Scottsdale, AZ, USA URL: https://proceedings.neurips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf
33. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics:

Human Language Technologies. 2019 Presented at: NAACL '19; June 2-7, 2019; Minneapolis, Minnesota p. 4171-4186 URL: https://aclanthology.org/N19-1423.pdf
34. Ji B, Li S, Yu J, Liu R, Tang JT, Yu J, et al. A BiLSTM-CRF method to chinese electronic medical record named entity recognition. In: Proceedings of the 2018 International Conference on Algorithms, Computing and Artificial Intelligence. 2018 Presented at: ACAI '18; December 21-23, 2018; Sanya, China p. 1-6 URL: https://dl.acm.org/doi/proceedings/10.1145/3302425 [doi: 10.1145/3302425.3302465]
35. Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging. arXiv. Preprint posted online on August 9, 2015 2023. [doi: 10.48550/arXiv.1508.01991]
36. Gridach M. Character-level neural network for biomedical named entity recognition. J Biomed Inform 2017 Jun;70:85-91 [FREE Full text] [doi: 10.1016/j.jbi.2017.05.002] [Medline: 28502909]
37. Wei H, Gao M, Zhou A, Chen F, Qu W, Wang C, et al. Named entity recognition from biomedical texts using a fusion attention-based BiLSTM-CRF. IEEE Access 2019 Jun 04;7:73627-73636 [FREE Full text] [doi: 10.1109/access.2019.2920734]
38. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. In: Proceedings of the 6th International Conference on Neural Information Processing Systems. 2013 Presented at: NIPS'13; December 5-10, 2013; Lake Tahoe, NV, USA p. 3111-3119 URL: https://dl.acm.org/doi/10.5555/2999792.2999959
39. Si Y, Wang J, Xu H, Roberts K. Enhancing clinical concept extraction with contextual embeddings. J Am Med Inform Assoc 2019 Nov 01;26(11):1297-1304 [FREE Full text] [doi: 10.1093/jamia/ocz096] [Medline: 31265066]
40. Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, et al. Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2018 Presented at: NAACL '18; June 1-6, 2018; New Orleans, LA, USA p. 2227-2237 URL: https://aclanthology.org/N18-1202.pdf [doi: 10.18653/v1/n18-1202]
41. Li N, Luo L, Ding Z, Song Y, Yang Z, Lin H. Improving Chinese clinical named entity recognition using stroke ELMo and transfer learning. In: Proceedings of the 2019 Evaluation Tasks at the China Conference on Knowledge Graph and Semantic Computing. 2019 Presented at: CCKS-Tasks '19; August 24-27, 2019; Hangzhou, China. [doi: 10.1007/978-981-15-1956-7_14]
42. Tutubalina E, Nikolenko S. Combination of deep recurrent neural networks and conditional random fields for extracting adverse drug reactions from user reviews. J Healthc Eng 2017;2017:9451342 [FREE Full text] [doi: 10.1155/2017/9451342] [Medline: 29177027]
43. Giorgi J, Bader GD. Towards reliable named entity recognition in the biomedical domain. Bioinformatics 2020 Jan 01;36(1):280-286 [FREE Full text] [doi: 10.1093/bioinformatics/btz504] [Medline: 31218364]
44. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020 Feb 15;36(4):1234-1240 [FREE Full text] [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
45. Kim YM, Lee TH. Korean clinical entity recognition from diagnosis text using BERT. BMC Med Inform Decis Mak 2020 Sep 30;20(Suppl 7):242 [FREE Full text] [doi: 10.1186/s12911-020-01241-8] [Medline: 32998724]
46. Wang Y, Ananiadou S, Tsujii J. Improving clinical named entity recognition in Chinese using the graphical and phonetic feature. BMC Med Inform Decis Mak 2019 Dec 23;19(Suppl 7):273 [FREE Full text] [doi: 10.1186/s12911-019-0980-z] [Medline: 31865903]
47. Yin M, Mou C, Xiong K, Ren J. Chinese clinical named entity recognition with radical-level feature and self-attention mechanism. J Biomed Inform 2019 Oct;98:103289 [FREE Full text] [doi: 10.1016/j.jbi.2019.103289] [Medline: 31541715]
48. Li Y, Du G, Xiang Y, Li S, Ma L, Shao D, et al. Towards Chinese clinical named entity recognition by dynamic embedding using domain-specific knowledge. J Biomed Inform 2020 Jun;106:103435 [FREE Full text] [doi: 10.1016/j.jbi.2020.103435] [Medline: 32360988]
49. Dong C, Zhang J, Zong C, Hattori M, Di H. Character-based LSTM-CRF with radical-level features for chinese named entity recognition. In: Proceedings of the 5th CCF Conference on Natural Language Processing and Chinese Computing, and 24th International Conference on Computer Processing of Oriental Languages. 2016 Presented at: NLPCC/ICCPOL '16; December 2-6, 2016; Kunming, China p. 2016 URL: https://link.springer.com/chapter/10.1007/978-3-319-50496-4_20 [doi: 10.1007/978-3-319-50496-4_20]
50. Qiu J, Wang Q, Zhou Y, Ruan T, Gao J. Fast and accurate recognition of Chinese clinical named entities with residual dilated convolutions. In: Proceedings of the 2018 IEEE International Conference on Bioinformatics and Biomedicine. 2018 Presented at: BIBM '18; December 3-6, 2018; Madrid, Spain p. 935-942. [doi: 10.1109/bibm.2018.8621360]
51. Yang X, Zou L. A conditional random fields approach to clinical name entity recognition. In: Proceedings of the Evaluation Tasks at the China Conference on Knowledge Graph and Semantic Computing. 2018 Presented at: CCKS '18; August 14-17, 2018; Tianjin, China p. 1-6 URL: https://ceur-ws.org/Vol-2242/paper01.pdf
52. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: a robustly optimized BERT pretraining approach. arXiv. Preprint posted online on July 26, 2019 2023. [doi: 10.48550/arXiv.1907.11692]

53. Cui Y, Che W, Liu T, Qin B, Yang Z. Pre-training with whole word masking for Chinese BERT. IEEE/ACM Trans Audio Speech Lang Process 2021;29:3504-3514 [FREE Full text] [doi: 10.1109/taslp.2021.3124365]
54. Crane K, Weischedel C, Wardetzky M. The heat method for distance computation. Commun ACM 2017 Oct 24;60(11):90-99 [FREE Full text] [doi: 10.1145/3131280]
55. Yidu-S4k: Yiducloud structured 4K dataset. OpenKG. URL: http://openkg.cn/dataset/yidu-s4k [accessed 2019-06-26]
56. Electronic medical records Center. Iiyi. URL: https://bingli.iiyi.com/ [accessed 2022-08-17]
57. Python software foundation. Python. URL: https://www.python.org/ [accessed 2022-10-28]
58. The Linux Foundation. Linux. URL: https://pytorch.org/ [accessed 2022-10-28]
59. ZDIC. URL: https://www.zdic.net/ [accessed 2023-04-17]
60. Jawahar G, Sagot B, Seddah D. What does BERT learn about the structure of language? In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019 Presented at: ACL '19; July 28- August 10, 2019; Florence, Italy p. 3651-3657 URL: https://aclanthology.org/P19-1356.pdf [doi: 10.18653/v1/p19-1356]
61. Zhang YQ, Wang Y. Identifying named entities of Chinese electronic medical records based on RoBERTa-wwm dynamic fusion model. Data Anal Knowl Discov 2022;6(2/3):242-250 [FREE Full text] [doi: 10.11925/infotech.2096-3467.2021.0951]
62. You XD, Ge HJ, Han JM, Li YX, Lü XQ. Recognition of complex entities in weapons and equipment field. Acta Sci Natur Univ Pekinensis 2022;58(3):391-404 [FREE Full text] [doi: 10.13209/j.0479-8023.2021.118]
63. Packard JL, Boltz WG. The origin and early development of the Chinese writing system. Language 1996 Dec;72(4):801 [FREE Full text] [doi: 10.2307/416104]
64. Cui SG, Chen JH, Xiaohong L. Named entity recognition of Chinese electronic medical records based on a hybrid neural network and medical MC-BERT. J Univ Sci Technol China 2022;51(04):565-571 [FREE Full text]
65. Meng Y, Wu W, Wang F, Li X, Nie P, Yin F, et al. Glyce: glyph-vectors for Chinese character representations. In: Proceedings of the 33rd Conference on Neural Information Processing Systems. 2019 Presented at: NeurIPS '19; December 8-14, 2019; Vancouver, Canada URL: https://papers.nips.cc/paper_files/paper/2019/hash/452bf208bf901322968557227b8f6efe-Abstract.html
66. Graves A, Ndez S, Schmidhuber J. Bidirectional LSTM networks for improved phoneme classification and recognition. In: Proceedings of the 15th International Conference on Artificial Neural Networks: Formal Models and Their Applications. 2005 Presented at: ICANN '05; September 11-15, 2005; Warsaw, Poland p. 799-804 URL: https://link.springer.com/chapter/10.1007/11550907_126 [doi: 10.1007/11550907_126]
67. Lafferty JD, McCallum AK, Pereira FC. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of the 18th International Conference on Machine Learning. 2001 Presented at: ICML '01; June 28- July 1, 2001; San Francisco, CA, USA p. 282-289 URL: https://dl.acm.org/doi/10.5555/645530.655813
68. Viterbi A. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Trans Inform Theory 1967 Apr;13(2):260-269 [FREE Full text] [doi: 10.1109/tit.1967.1054010]
69. Kong J, Zhang L, Jiang M, Liu T. Incorporating multi-level CNN and attention mechanism for Chinese clinical named entity recognition. J Biomed Inform 2021 Apr;116:103737 [FREE Full text] [doi: 10.1016/j.jbi.2021.103737] [Medline: 33737207]
70. Li Y, Wang X, Hui L, Zou L, Li H, Xu L, et al. Chinese clinical named entity recognition in electronic medical records: development of a lattice long short-term memory model with contextualized character representations. JMIR Med Inform 2020 Sep 04;8(9):e19848 [FREE Full text] [doi: 10.2196/19848] [Medline: 32885786]
71. Wan Q, Liu J, Wei L, Ji B. A self-attention based neural architecture for Chinese medical named entity recognition. Math Biosci Eng 2020 May 09;17(4):3498-3511 [FREE Full text] [doi: 10.3934/mbe.2020197] [Medline: 32987540]
72. Wang C, Wang H, Zhuang H, Li W, Han S, Zhang H, et al. Chinese medical named entity recognition based on multi-granularity semantic dictionary and multimodal tree. J Biomed Inform 2020 Nov;111:103583 [FREE Full text] [doi: 10.1016/j.jbi.2020.103583] [Medline: 33010427]

Abbreviations

BERT: Bidirectional Encoder Representation from Transformers
BiLSTM: Bidirectional Long Short-Term Memory
CNER: clinical named entity recognition
CNN: convolutional neural network
CRF: conditional random fields
ELMo: Embeddings from Language Models
EMR: electronic medical record
FN: false negative
FP: false positive
LSTM: long short-term memory
NER: named entity recognition
PLM: pretrained language model

RoBERTa: Robustly Optimized Bidirectional Encoder Representation from Transformers Pretraining Approach
TP: true positive
Word2Vec: Word to Vector
wwm: Whole Word Masking

Edited by G Eysenbach; submitted 01.12.22; peer-reviewed by C Yu, L Heryawan; comments to author 31.01.23; revised version received 18.02.23; accepted 31.03.23; published 10.05.23.

Please cite as:
Wang W, Li X, Ren H, Gao D, Fang A
Chinese Clinical Named Entity Recognition From Electronic Medical Records Based on Multisemantic Features by Using Robustly Optimized Bidirectional Encoder Representation From Transformers Pretraining Approach Whole Word Masking and Convolutional Neural Networks: Model Development and Validation
JMIR Med Inform 2023;11:e44597
URL: https://medinform.jmir.org/2023/1/e44597
doi:10.2196/44597
PMID:37163343

Original Paper

Improving an Electronic Health Record–Based Clinical Prediction Model Under Label Deficiency: Network-Based Generative Adversarial Semisupervised Approach

Runze Li[1*], BE; Yu Tian[1*], PhD; Zhuyi Shen[1], BE; Jin Li[2], PhD; Jun Li[3], MD; Kefeng Ding[3], MD; Jingsong Li[1], PhD

[1]College of Biomedical Engineering and Instrument Science, Zhejiang University, Hangzhou, China

[2]Institute for Artificial Intelligence in Medicine, School of Artificial Intelligence, Nanjing University of Information Science and Technology, Nanjing, China

[3]Department of Surgical Oncology, Zhejiang University School of Medicine Second Affiliated Hospital, Hangzhou, China

[*]these authors contributed equally

Corresponding Author:
Jingsong Li, PhD
College of Biomedical Engineering and Instrument Science
Zhejiang University
Zhou Yiqing Science and Technology Building, 2nd floor
38 Zheda Road
Hangzhou, 310027
China
Phone: 86 571 87951564
Fax: 86 571 87951564
Email: ljs@zju.edu.cn

Abstract

Background: Observational biomedical studies facilitate a new strategy for large-scale electronic health record (EHR) utilization to support precision medicine. However, data label inaccessibility is an increasingly important issue in clinical prediction, despite the use of synthetic and semisupervised learning from data. Little research has aimed to uncover the underlying graphical structure of EHRs.

Objective: A network-based generative adversarial semisupervised method is proposed. The objective is to train clinical prediction models on label-deficient EHRs to achieve comparable learning performance to supervised methods.

Methods: Three public data sets and one colorectal cancer data set gathered from the Second Affiliated Hospital of Zhejiang University were selected as benchmarks. The proposed models were trained on 5% to 25% labeled data and evaluated on classification metrics against conventional semisupervised and supervised methods. The data quality, model security, and memory scalability were also evaluated.

Results: The proposed method for semisupervised classification outperforms related semisupervised methods under the same setup, with the average area under the receiver operating characteristics curve (AUC) reaching 0.945, 0.673, 0.611, and 0.588 for the four data sets, respectively, followed by graph-based semisupervised learning (0.450, 0.454, 0.425, and 0.5676, respectively) and label propagation (0.475,0.344, 0.440, and 0.477, respectively). The average classification AUCs with 10% labeled data were 0.929, 0.719, 0.652, and 0.650, respectively, comparable to that of the supervised learning methods logistic regression (0.601, 0.670, 0.731, and 0.710, respectively), support vector machines (0.733, 0.720, 0.720, and 0.721, respectively), and random forests (0.982, 0.750, 0.758, and 0.740, respectively). The concerns regarding the secondary use of data and data security are alleviated by realistic data synthesis and robust privacy preservation.

Conclusions: Training clinical prediction models on label-deficient EHRs is indispensable in data-driven research. The proposed method has great potential to exploit the intrinsic structure of EHRs and achieve comparable learning performance to supervised methods.

(JMIR Med Inform 2023;11:e47862) doi:10.2196/47862

KEYWORDS

semisupervised learning; generative adversarial network; network analysis; label deficiency; clinical prediction; electronic health record; EHR; clinical prediction; adversarial network; data set

Introduction

The recent rise of observational biomedical research, driven by greatly expanding electronic health records (EHRs) and the prevalence of machine learning methods, has drawn great attention [1-4]. Conventional strategies tend to screen out subgroups of interest based on expert supervision or established risk factors. An alternative data-driven paradigm extracts underlying subtypes by comprehensively profiling the longitudinal irregularity, interdimensional heterogeneity, and intrinsic homogeneity of the database, thus progressively facilitating the practice of precision medicine. For instance, the Electronic Medical Records and Genomics (eMERGE) network [5] leverages expertise from multiple institutions and communities to integrate biorepositories and EHRs to support genomic research. Observational research approaches exhibit both potential and challenges for more sophisticated data analysis.

However, the acquisition of realistic data, especially data labels, is still restricted when confronting concerns about system security, patient privacy, and intellectual property protection [6,7]. Excluding data and labels may be ubiquitous during the data collection phase. Long-term studies often lack sufficient time to gather data and have no control over the switching behaviors of patients [8,9], resulting in the loss of accurate outcome measurements.

Restrictions on transferring intellectual property among different institutions hinder the sharing of data, which is expected to be complete. Additionally, some expertise-requiring annotations are tedious and have no guarantee of correctness [10]. Generally, label deficiencies occur frequently when analyzing observational EHR data.

There have been some attempts to address insufficient labeling by realistic synthesized EHR (RS-EHR) generation. One approach to RS-EHRs is knowledge-based [11,12]. Such approaches combine publicly available statistics, clinical practice guidelines, and medical coding dictionaries to improve the fidelity of generated EHRs. However, the models are still restricted to development, testing, and public demonstrations.

Another strategy is data-driven. Generative adversarial networks (GANs) are a new class of methods for obtaining realistic synthesized data [13,14]. The philosophy of GANs is to train two networks, one generating fake samples and the other discriminating fake and real samples, in a min-max game until equilibrium is achieved, indicating that the generated fake samples cannot be distinguished from the real samples. There has been some work on applying state-of-the-art GANs to generate synthesized EHR data sets [15,16]. However, these studies have not fully applied the generated data to augment EHR computational phenotyping and classification. GANs for few-labeled data are still unlikely to recover the whole distribution of labels from the raw data set due to imbalanced labeling. Additionally, there are some arguments that GAN-generated samples are likely to copy real samples exactly, which is a potential violation of privacy [17,18].

Semisupervised learning (SSL) is a set of techniques that are usually adopted to leverage unlabeled data and an underlying data set structure. With a relatively small set of labeled data compared to that needed in supervised learning (SL), SSL can still display decent learning performance. Some previous studies used SSL to phenotype EHR databases [19,20]. These studies achieved excellent performance in EHR-based risk prediction, but the feature dimensions were restricted to discrete medical codes. GANs were adopted to boost the SSL [21], but as mentioned above, the generator was trained to eventually remember exact copies of the samples for the limited span of an EHR data set in a discrete and high-dimensional space, which therefore raised privacy concerns. SSL is a powerful tool for label-deficient circumstances but needs specifications for observational research.

Network analysis is a solution to both obstacles. Encoding the similarities among patients into their connections protects their identities. The input of the analysis is only the network structure and embedded vectors. Network analysis is the basis for manifold learning, which has an advantage in approximating the data structure in a high-dimensional space. Many manifold-based methods have prevailed in intuitively visualizing and phenotyping coordinated data sets [22-24]. Additionally, there have been quite a few attempts to extend deep learning to irregular data structures, such as graphs. Several studies have shown great performance in representative learning with SSL [25-28], and endeavors have been made to apply GANs to graphs [29-31].

However, few studies have considered exploiting the inherent network structure of an EHR database in SSL tasks. GANs on networks have not been fully investigated in terms of privacy preservation. Additionally, under various label-deficient situations, the performance remains to be evaluated. It is very promising to scale SSL and GANs to the graph structure extracted from an EHR database and to thereby acquire a new perspective on EHR data sets.

For this paper, the main contributions are as follows: (1) This study tries to address limitations due to existing label deficiency in observational EHR analytical research by extending the network analysis pipeline to EHRs. A boosting learning model is proposed by applying GAN-boosted SSL to network data extracted from label-deficient coordinated EHRs. (2) Experiments are conducted on 3 public data sets as well as one from the First Affiliated Hospital of Zhejiang University, and they are evaluated by prediction metrics that are compared to conventional learning methods. The proposed method shows superior performance over conventional semisupervised methods and indicates comparable performance with supervised learning methods when data are fully labeled. (3) To ensure the utility of the proposed model, further evaluations of data quality, nondisclosure, and memory space consumption are performed.

The proposed method shows higher data fidelity, lower precision metrics against compromised attack, and less graphics processing unit (GPU) memory consumption over conventional semisupervised methods.

Methods

Data Set Structure Conversion to a Graph

Graph structure definition and semisupervised learning on graph formularization are shown in Multimedia Appendix 1A [31-37]. According to the well-accepted assumption that a manifold is locally Euclidean in topological space, it is plausible to represent a data set X with a graph G. However, this conversion rule should be scrutinized. First, it depends on the number of edges |E| that comprise the edge set. |E| should be restricted to a range that avoids disconnected components and short circuits that obscure structural information. Second, the neighborhood searching strategy should be scalable to feature value scales and effective in practice. Third, the local density variance should be preserved during conversion, which means that the weights of edges should not be binary.

To circumvent this problem, the k-nearest neighbors (k-NN) method was selected to convert the original data set into a graph measure space. As its name indicates, the k nearest points in the Euclidean space of point x are identified as its neighbors, N_k (x). Each edge weight w_{ij} is refined with the Jaccard coefficient:

The Jaccard coefficient addresses the unified weight problem brought by k-NN searching and restricts the weights to [0,1], which scales the local densities as node degrees: $\deg(v_i) = \sum_{j \in N\_k(v\_i)} w_{ij}$. Additionally, when the lower bound is reached, the edge is removed from the graph, and eventually, nodes with degree zero will be considered noise and therefore removed. The final graph serves as one of the inputs of the GAN.

GANs for Graphs and Their Modified Losses

In this study, we focus on generating vectorized fake samples by the use of both the graph structure and coordinated features. The coordinated features of the graph structure are acquired by feeding the Jaccard graph into large-scale information network embedding [38], explicitly setting the output dimension as half of d.

The fundamentals of GAN [32,33] are presented in Multimedia Appendix 1C [31-37]. Nevertheless, it is important to take into account the unconventional loss of semisupervised adversarial learning, as insufficient labels do not effectively minimize the current adversarial learning loss. The generator is trained to produce samples that bridge the density gap between samples from distinct classes. In the case of binary classification tasks, the 2 classes are “true” and “false.” By expanding the density gap between labeled true samples, labeled false samples, and generated density gap samples, the adjusted discriminator loss can enhance semisupervised learning performance. The refined discriminator loss L_D for SSL purposes comprises semisupervised loss, entropy loss, and class distance. (1) Semisupervised loss: there are two terms; the first is the supervised loss calculated by cross-entropy between the labels and prediction. The second emphasizes the loss due to incorrect classification by SSL. λ_0 is a hyperparameter that balances these 2 terms.

$$\begin{aligned} loss_{semi} &= loss_{sup} + \lambda_0 loss_{un} \\ &= -E\_(_{xi \in XL}) \log P(y_i \mid x_i, y_i < M) \\ &- \lambda_0 \left(E_{xi \in XU} \log(1 - P(M \mid x_i)) + E_{xi \sim G(z)} \log P(M \mid x_i)\right) \end{aligned} \quad \textbf{(2)}$$

(2) Entropy regularization [39,40]: this term calculates the entropy of a distribution over all labels M to enhance the certainty of the prediction.

(3) Cluster distance loss [31]: this term tends to enlarge the density gap so that samples from different classes are separate. $h^n(x)$ is the last-layer output of the discriminator.

The final loss term for the discriminator generator is

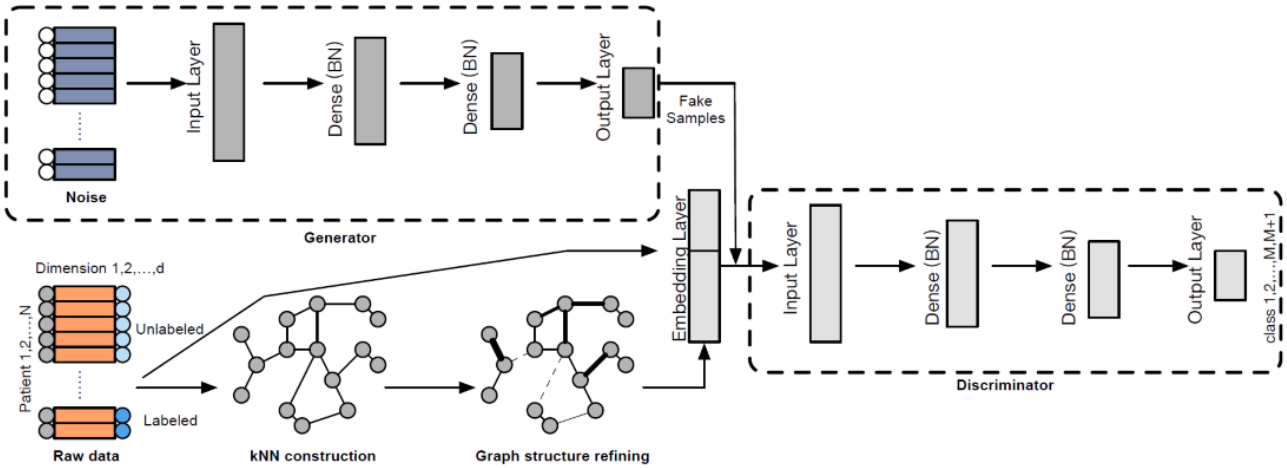
$$loss_D = loss_{Dwgp} + loss_{semi} + loss_{ent} + loss_{pt} \quad \textbf{(5)}$$

The loss of generator $_{LG}$ is also modified by adding the term (4). The final loss term for the training generator is

$$loss_G = loss_{Gwgp} + loss_{pt} \quad \textbf{(6)}$$

The network structure is illustrated in Figure 1. During the training phase, real embeddings and fake inputs of the same size are fed into the network separately in batches, with the goal of optimizing the aforementioned losses of the discriminator and generator. During 1 training epoch, batches of real labeled data, real unlabeled data, and fake data are fed into the network to calculate different loss terms for optimization. Batch normalization is conducted. After training, the discriminator loss is expected to be stable and could be exploited as a classifier for testing samples and predictions. The generator is suitable for measuring data quality and preserving privacy.

Figure 1. An overview of the proposed model. Real samples are extracted from the k-NN graph by the embedding layer, and the fake samples generated by the generator have the same dimension. Both real and fake samples are fed into and backpropagated from the discriminator in minibatches. The output layer of the discriminator is softmax. The embedding layer is a pretrained large-scale information network embedding [32]. BN: batch normalization; k-NN: k-nearest neighbors.



Data Sets and Experimental Setup

EHR data sets were obtained from public resources, including University of California Irvine Machine Learning Repository Type 2 Diabetes 30-Day Readmission (UCI-T2D) [41]; Surveillance, Epidemiology, and End Results Ovarian Cancer (SEER-OVC) [42]; and Surveillance, Epidemiology, and End Results Colorectal Cancer (SEER-CRC) [42]. The dimensional information is summarized in Table 1. Another colorectal cancer data set from the Second Affiliated Hospital Zhejiang University School of Medicine (SAHZU-CRC) was selected to investigate feasibility in practical situations. These data sets were selected because they are long-term follow-ups, the labels of which take much time and effort to obtain and are likely to be missing due to regulations on data collection. The selected features included basic demographics, medication, clinical codes, stage codes, laboratory variables, and dispositions. A basic description of the data sets and preprocessing is provided in Multimedia Appendix 1B.

We trained the proposed models for a maximum of 200 epochs using Adam optimization with a learning rate of 0.003 and a momentum of 0.5. The batch size was 128. For each class, the rate of labeled points (the label rate is the percentage of labeled points among all points) increased progressively from 5% to 25% with a step of 5%. The number of test sets was set as 20% of the data set.

Table 1. Dimensional description of the selected data sets.

Data sets	Records	Categorical variables	Numerical variables	Preprocessed dimensions	Labeling standard
University of California Irvine Machine Learning Repository Type 2 Diabetes 30-Day Readmission	61,675	44	8	57	Readmission in 30 days
Surveillance, Epidemiology, and End Results Ovarian Cancer	10,038	18	3	34	Survival over 5 years
Surveillance, Epidemiology, and End Results Colorectal Cancer	40,014	7	2	14	Survival over 5 years
Second Affiliated Hospital of Zhejiang University	1244	8	2	14	Survival over 5 years

We compared the model with the following baselines: (1) supervised learning methods, including logistic regression (LR), a support vector machine (SVM), and a random forest (RF) and (2) SSL methods, including graph-based semisupervised learning (GSSL) and label propagation (LP). All these methods are run using the *scikit-learn* Python package. The graph convolutional network (GCN) [25,27], a state-of-the-art graph-based semisupervised deep learning method, is also considered a competing method. To measure the classification performance, the accuracy and recall—for the important purpose of excluding false negative cases to conserve medical resources—and the area under the receiver operating characteristics curve (AUC) were selected as metrics. Each metric represented the average of 30 repetitions of 10-fold cross-validation training.

Ethical Considerations

This study did not involve any human or animal experiments. The UCI-T2D, SEER-OVC, and SEER-CRC data sets are public, and we complied with their ethical requirements. We also used a colorectal cancer–specific disease cohort of the Second Affiliated Hospital Zhejiang University School of Medicine; this was approved by the Human Research Ethics Committee of Zhejiang University in August 2017 (2017-067).

Results

SSL-based Classification of EHR Data

In the aforementioned experiments, the proposed method for semisupervised classification outperformed related semisupervised methods by a decent margin. Basic graph

semisupervised methods (ie, GSSL) are limited in classification performance, mostly due to their assumption that edges encode only the similarity of nodes. The spectral methods (LP and GCNs) do not perform well, perhaps because their low-order approximation may smooth the frontiers in the graph. Neither of these 2 methods consider the local properties of the input graph, and under some circumstances, they classify the majority of nodes into 1 class. Additionally, at a 10% to 15% label rate, the proposed method achieves the best performance on SEER-OVC, SEER-CRC, and SAHZU-CRC (Table 2). The AUCs declined as label rates continued to rise. GCNs, as the state-of-the-art semisupervised deep learning method, had somewhat better results for a data set with a size that can be handled by a GPU, but still exhibited worse performance than the proposed method, presumably due to oversmoothing the graphs and having less refined loss.

In regard to supervised learning, as shown by the bars in Figure 2A, even with a label rate of 10%, SSL on a graph with a GAN performed comparably to the supervised learning methods. As the portion of labeled data increased, the learning performance progressively increased, which is a consequence of the abundant information of the label distribution over the constructed graph. However, as the label rate continued to rise, the performance decreased because of mode collapse and overfitting. As the error bars show, with 10% labeled data, the standard deviations of the proposed model are slightly larger, as shown in Figures 2A and 2C, indicating a limitation of our proposed method; it only applies to certain low label-rate circumstances. When the labels are sufficient, more robust SL methods are better. However, some poorly trained and undertuned SL methods show far worse metrics in testing. Additionally, as the vector dimensions, shown in Table 1, decreased somewhat, the learning performance showed a significant decrease. This is perhaps a consequence of the lack of dimension diversity for similarity encoding and the local graph structure.

Table 2. Summary of the results of the classification AUCs for semisupervised learning methods under progressively increasing label rates. The learning performance of the graph convolutional network on the large data sets—that is, data sets other than Second Affiliated Hospital of Zhejiang University Colorectal Cancer—is unavailable due to memory limits.

Label Rate	5%, AUC[a]	10%, AUC	15%, AUC	20%, AUC	25%, AUC
University of California Irvine Machine Learning Repository Type 2 Diabetes 30-Day Readmission					
GSSL[b]	0.450	0.472	0.523	0.542	0.602
LP[c]	0.475	0.475	0.564	0.585	0.566
Proposed	0.929	0.979	0.964	0.930	0.924
Surveillance, Epidemiology, and End Results Ovarian Cancer					
GSSL	0.454	0.512	0.537	0.591	0.591
LP	0.344	0.364	0.462	0.478	0.491
Proposed	0.640	0.719	0.677	0.678	0.650
Surveillance, Epidemiology, and End Results Colorectal Cancer					
GSSL	0.525	0.527	0.447	0.585	0.578
LP	0.540	0.532	0.512	0.540	0.513
Proposed	0.595	0.652	0.640	0.581	0.590
Second Affiliated Hospital of Zhejiang University Colorectal Cancer					
GSSL	0.547	0.573	0.564	0.553	0.580
LP	0.454	0.448	0.512	0.460	0.507
GCN[d]	0.505	0.575	0.562	0.585	0.606
Proposed	0.587	0.650	0.634	0.568	0.508

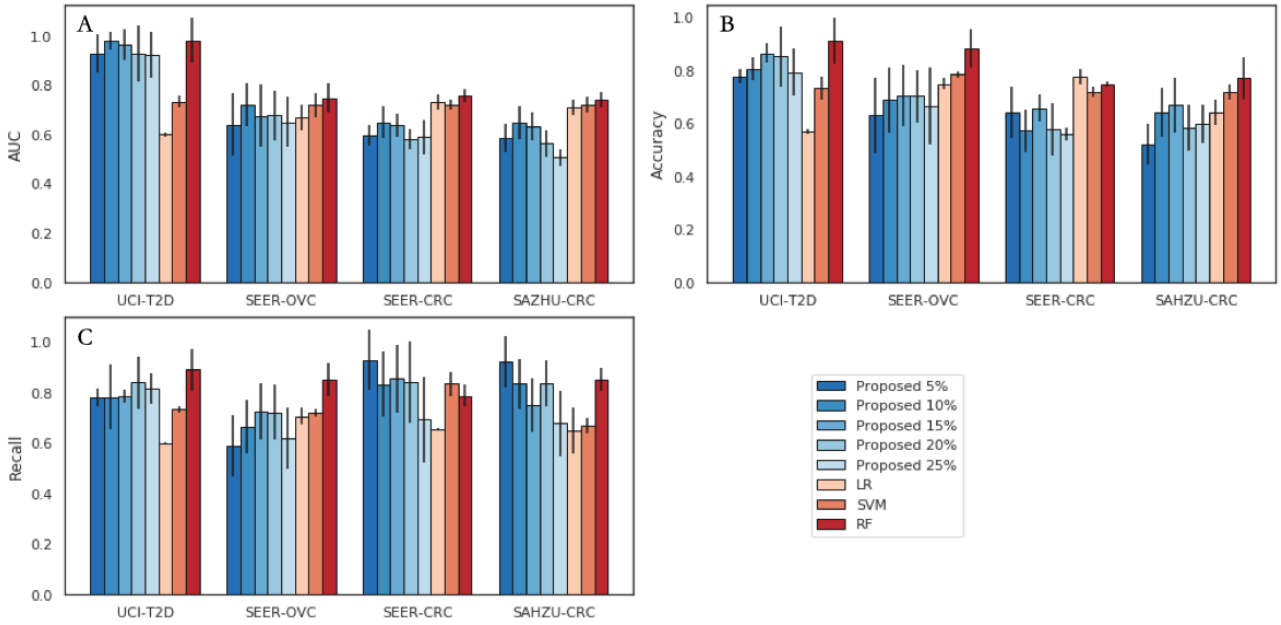
[a]AUC: area under the receiver operating characteristics curve.

[b]GSSL: graph-base semisupervised learning.

[c]LP: label propagation.

[d]GCN: graph convolutional network.

Figure 2. Summary of the results of the classification of the proposed method versus those of the conventional supervised learning methods. (A) AUC; (B) accuracy; (C) recall. The proposed method was evaluated under progressively increasing label rates. The supervised learning models were trained on fully labeled data. The error bars indicate the SD for each metric. AUC: area under the receiver operating characteristics curve; LR: logistic regression; RF: random forest; SAHZU-CRC: Second Affiliated Hospital of Zhejiang University Colorectal Cancer; SEER-CRC: Surveillance, Epidemiology, and End Results Colorectal Cancer; SEER-OVC: Surveillance, Epidemiology, and End Results Ovarian Cancer; SVM: support vector machine; UCI-T2D: University of California Irvine Machine Learning Repository Type 2 Diabetes 30-Day Readmission.

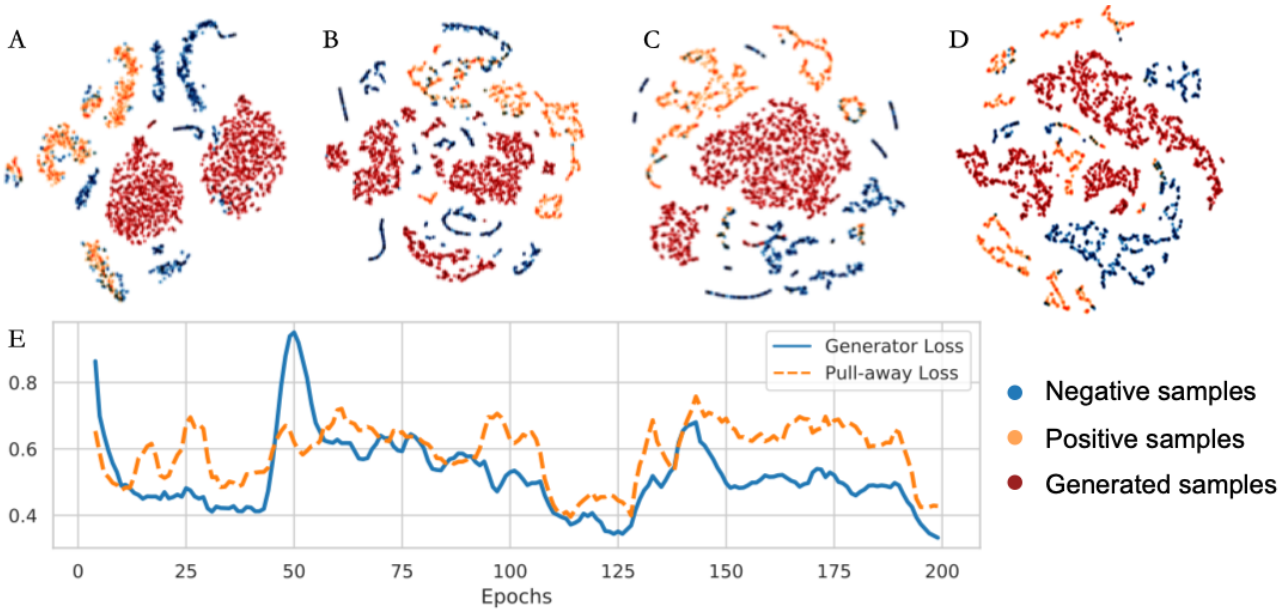


Boosting Semisupervised Learning by Generating a Density Gap

In this section, we visualize the final layer of discriminator D in the proposed method by feeding it real samples from UCI-T2D and their generated counterparts. By embedding the output layer at different iteration steps with t-distributed stochastic neighbor embedding [22], the progression of the density gap from the generated samples, described in equation 4, is verified.

In Figure 3, we can see that at the starting epochs, the generated samples are mixed with the real samples, and the different classes are not divided. During training, D gradually learns a nonlinear map to project the fake samples and real samples into distinct clusters, while G learns to generate samples to take over the central area and isolate the clusters of different classes. This process has 2 advantages. First, the fake samples from the generator are unlikely to be copies of the original data, avoiding the direct disclosure of private information. Second, the samples on the borders of different classes are more correctly divided, which not only improves the accuracy of classification but also reveals the underlying training strategy of splitting one large class into several smaller classes to obtain a better classification.

Figure 3. The progressive generation of density gaps in high-dimensional space and its visualization. (A) 0 epochs; (B) 40 epochs; (C) 80 epochs; (D) 120 epochs. The generated samples ultimately span the gap between a real and a false sample. The line chart (E) indicates the function of the pull-away term and how its optimization affects the generator and discriminator.



Fidelity Evaluation of the Generated Data

Frontier nodes are nodes at the borders of different clusters in a graph. The definition is given in Multimedia Appendix 1D. It is possible that a trained model is exploited directly for secondary purposes, such as fundamental profiling or developmental usage during the primary phase of data sharing [12]. We calculated the dimensionwise probability (DWPro) and dimensionwise prediction (DWPre) proposed by Choi et al [15] to evaluate the fidelity of the generator in our proposed model. DWPro is a basic statistical confirmation of the distributions of real data that are appropriately learned by the generator in the model. A training set R and synthetic sample set S of the same sample size are compared using the Bernoulli success probability pk of each dimension k. DWPre measures the extent to which the internal relations of every feature are captured. One dimension k is selected, and the rest of the features are used as training data. R and S are used to train the LR classifiers. Then, the dimension k is regarded as the label column for testing. It is a rational assumption that a smaller margin between the predictions of 2 models implies a better synthetic quality. The F_1-score is selected as the metric for comparison.

Figure 4 shows that all 4 data sets were depicted well from a featurewise perspective, and over half of the dots are near the diagonal line. In Figure 4C, the consistency of each feature indicates high synthetic quality. Figure 5 shows a mildly diminished learning quality considering interdimensional fidelity. However, half of the features are still likely to be inferred from the remaining columns and the same proportion of features. Considering that the generated frontier is still different from the directly generated datapoints [15,16], the fidelity is acceptable for some secondary uses.

Figure 4. Dimensionwise probability of 4 selected data sets: (A) University of California Irvine Machine Learning Repository Type 2 Diabetes 30-Day Readmission; (B) Surveillance, Epidemiology, and End Results Ovarian Cancer; (C) Surveillance, Epidemiology, and End Results Colorectal Cancer; (D) Second Affiliated Hospital of Zhejiang University Colorectal Cancer. The x-axis is the Bernoulli success probability for the features of the real data sets, while the y-axis is the corresponding value from the synthetic data. Each blue dot represents a feature of the data set. The red diagonal line indicates an identical Bernoulli success probability of both the real and generated data sets, and ideal fidelity is learned featurewise by the generator.

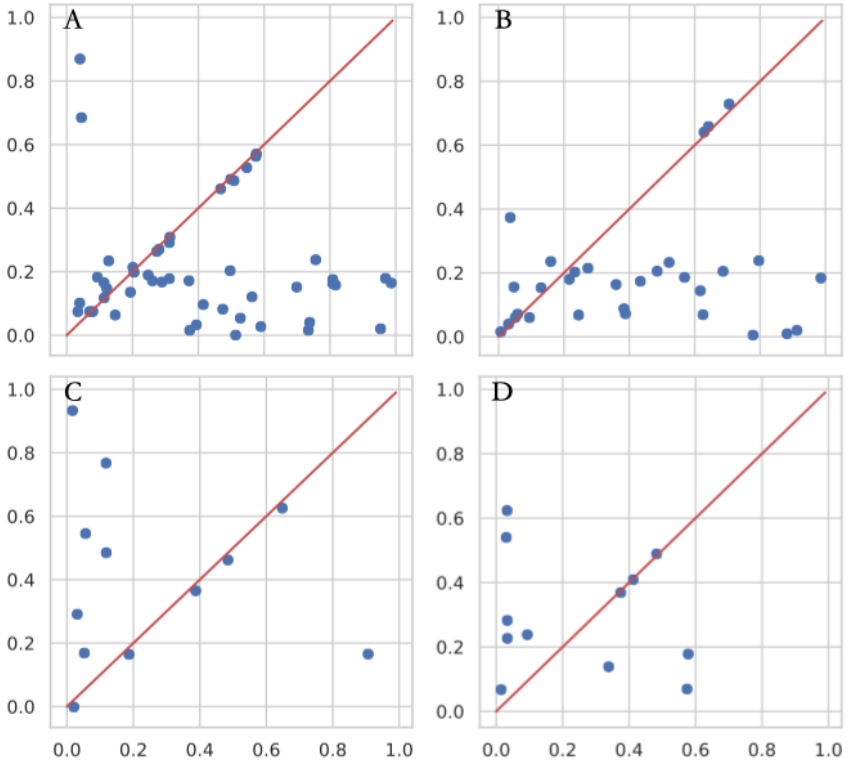
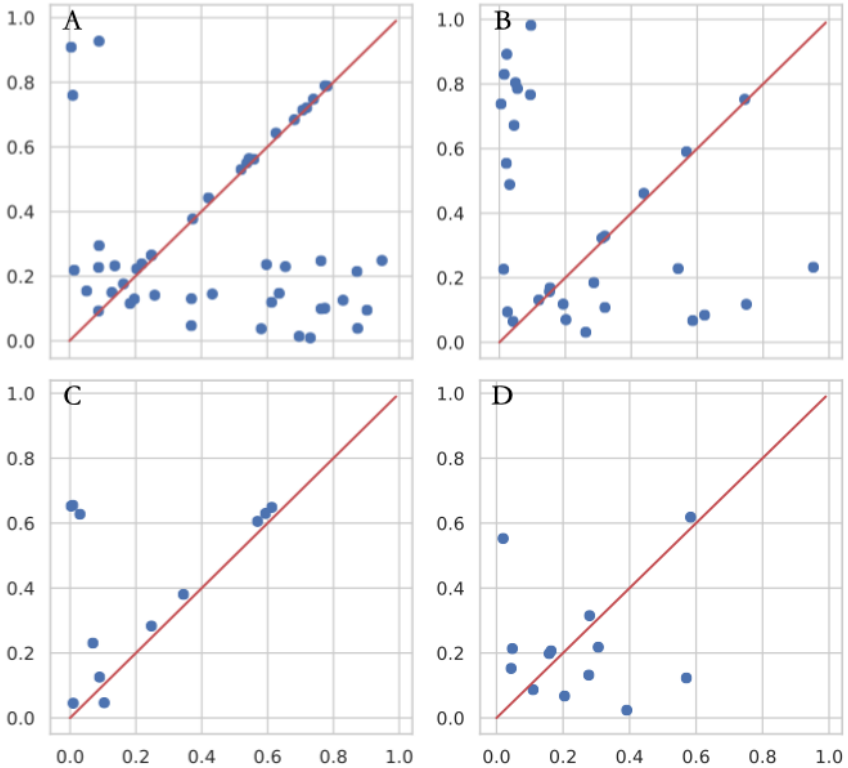


Figure 5. Dimensionwise prediction of 4 selected data sets: (A) University of California Irvine Machine Learning Repository Type 2 Diabetes 30-Day Readmission; (B) Surveillance, Epidemiology, and End Results Ovarian Cancer; (C) Surveillance, Epidemiology, and End Results Colorectal Cancer; (D) Second Affiliated Hospital of Zhejiang University Colorectal Cancer. The x-axis is the F_1-score of models trained on the real data sets, while the y-axis is the corresponding values from the synthetic data. Each blue dot represents a feature of the data set. The red diagonal line indicates that the F_1-score was identical for the models trained and tested on the real and generated data sets, and ideal interdimensional fidelity was learned by the generator.

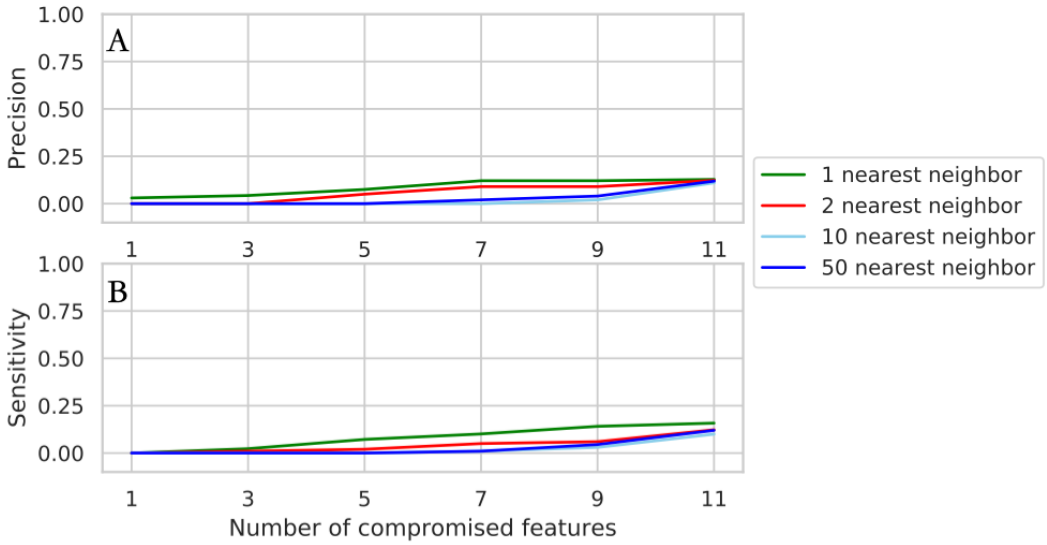


Evaluation of the Disclosure Risk of the Generated Data

The generator in our proposed model may be exploited to generate data points similar to the original data sets, posing threats to patient privacy. We need to ensure that the frontier nodes generated by the proposed model can be protected from attackers with compromised data. Therefore, a quantitative evaluation of presence and attribute disclosure was conducted on the SAHZU-CRC data set. Of the real samples N, 1% were randomly sampled, and among the 11 dimensions (the numerical dimensions are left out and the nominal columns are collapsed into 2 for simplification), a progressively increasing number of features, denoted as r, were assumed to be known by the attacker. Then, the attacker could exploit the knowledge of the data (1% × N × r) to conduct k-NN searches of the synthetic data, and the other unknown feature values were estimated according to those of the k-NN. Finally, the unknown features were compared to the real features to gain precision and accuracy. The calculation was repeated 100 times with 1% of the real records chosen at random.

Under these circumstances, the sensitivity indicates when the attacker has 1% × N × r of the disclosed data and all the synthetic data and how many records of all the positive features the attacker can correctly estimate with a 1-NN attack. The precision indicates how many features among all the features estimated positive by the attack were on average accurate. For instance, in Figure 6A, if an attacker with 1% of the records (12 of 1244 records for SAHZU-CRC) and 5 features from the real data conducts 1-NN on the synthetic SAHZU-CRC data generated by the proposed method, the positive estimation of the remaining unknown features of the real data will be 12.5% correct on average (0.125 precision), and of all the positive predictions, 15.8% will be correct (0.158 sensitivity).

Figure 6. Privacy preservation evaluation. When increasing the number of known features, the attack achieves (A) precision and (B) sensitivity with 1% compromised records from the Second Affiliated Hospital of Zhejiang University data set.



In summary, the precision and sensitivity of attacks on synthetic data is relatively low, 0.158 at best when r is 11. The most effective attack setting is 1-NN. It is difficult to estimate more information from our frontier nodes due to the modification of the network losses. Substitution of both the generator and discriminator learning strategies boosts the model performance on classification with label deficiency and provides synthetic samples capable of preventing disclosure.

Scalability of the Memory Consumption of Batch-Based Training

Because GPUs have been used in deep learning–based computation, we further examined whether the proposed method could achieve practical memory consumption. The aforementioned semisupervised methods are compared with our proposed method on memory consumption for 4 data sets. For the algorithms that do not need GPUs, their central processing unit consumption is measured.

For small data sets (eg, SAHZU-CRC), our proposed method takes up more space because of its complex network structure (Table 3). However, as the size expands, our proposed method shows the least and most stable space consumption, because minibatch training is independent of the number of samples (for SEER-CRC and SEER-OVC). Conventional network-based SSL methods tend to train on full batches. When the data set is large enough, there is a huge obstacle to storing the data in memory. Stable memory consumption implies a scalable model for training and prediction on diverse data sizes. The GCN, as a transductive SSL method, is unable to be directly scaled to larger data sets despite its excellent representative ability.

Table 3. Graphics processing unit memory consumption of the proposed method against that of typical semisupervised learning algorithms on 4 data sets. The graph convolutional network is not suitable for data sets other than Second Affiliated Hospital of Zhejiang University; therefore, only one result is shown.

	UCI-T2D[a], MB	SEER-OVC[b], MB	SEER-CRC[c], MB	SAHZU-CRC[d], MB
Graph-based semisupervised learning (CPU[e])	1374	770	1260	297
Label propagation (CPU)	1200	702	1263	257
Graph convolutional network	Out of memory	Out of memory	Out of memory	732
Proposed	336	345	330	332

[a]UCI-T2D: University of California Irvine Machine Learning Repository Type 2 Diabetes 30-Day Readmission.

[b]SEER-OVC: Surveillance, Epidemiology, and End Results–Ovarian Cancer.

[c]SEER-CRC: Surveillance, Epidemiology, and End Results–Colorectal Cancer.

[d]SAHZU-CRC: Second Affiliated Hospital of Zhejiang University.

[e]CPU: central processing unit.

Discussion

Principal Results

The proposed model fully utilizes the inner graphical structure of EHRs and provides cost-effective prediction metrics. The density gap derived from the modified network loss enables different class labels to be better distinguished. Under label-deficient circumstances, the proposed model achieves a comparable performance to that of conventional supervised learning methods where all of the training data are labeled. Specifically, with only 10% labeled data, the performance of popular supervised machine learning methods is approached, which implies there is a broad set of situations in which this model could be considered for prediction tasks. Following the same setting of label rates for the purpose of comparison, the conventional SSL methods show poor data representation ability. The learning performance, compared to that of our proposed method, shows worse stability and scalability. With the increasing label rate, the conventional SSL models display either poor performance on classification due to label deficiency or extreme cases where the classifier puts every sample into 1 class as a consequence of overfitting. Additionally, the memory cost is worth noting. Most semisupervised methods have a tendency to copy the whole graph structure into memory [19,27,43], which brings a very large burden of computational resources considering that the EHRs absorb increasingly more data.

Extracting the frontier of generated samples that shows high performance in DWPro and DWPre has potential in applying some special frontier nodes as sample data for secondary usage, in the same way as related work applies GANs to RS-EHR generation. According to related studies [15,16], generating data with adequate quality is crucial in cross-organization data sharing. The quality of the data determines the model performance on realistic data sets. Additionally, for diverse developmental needs, the more realistic the generated data are compared to the real samples, the more persuasiveness and fidelity the researching systems will acquire. The generator in our model fulfills this demand by generating similar samples to the original data after the training phase.

To reveal the hidden clinical and physiological characteristics of certain groups, EHRs are among the most reliable information sources. Nonetheless, administrative regulations and the protection of patient privacy have decreased the accessibility of EHRs for a variety of reasons and made downstream analysis inconvenient. Our method first addresses privacy considerations by transforming the data set into a k-NN graph where the similarities between different patients are re-encoded while the identifying information is hidden. Second, the vector from the embedded graph is fed into our model for further analysis. Under practical scenarios, authorization to share and use the original data will not be a necessity. Additionally, even when conventional attacks attempt to reidentify personal information from the publicly generated samples, the k-precision and k-sensitivity metrics indicate that it is quite safe if the attacker holds only a small fraction of the knowledge of the real data and conducts the most powerful 1-NN attack. Furthermore, the density gaps avoid the usual case where GANs would otherwise be trained to copy the real input, thereby shielding the patient information from another possible method of disclosure.

Limitations

The limitations of this study are still worth noting. The evaluation of how the proposed model can improve data quality and predict performance on the actual label collection phase has yet to be considered. Additionally, we excluded all patient duplicates to conduct a prediction method without considering any temporal information. Further investigation of the temporal trajectories of the same patients may reveal more of the inner mechanisms of disease progression, and localization methods of temporal and spatial structure in many other fields may address the same problem [44,45]. Additionally, the proposed model only applies to some label-rate setups, and performance diminishes as more labels become available. The thresholds for switching between the different algorithms (SSL and SL) remain to be studied. Finally, to be more protective of patient privacy and intellectual property, our future explorations include graph generation and attention mechanisms [28,29,34,46]. A whole generated graph can be taken into consideration. With the power of GANs, the underlying structure of large-scale EHRs could be preserved while achieving full anonymity.

Conclusions

EHR-based systems and observational studies with conventional learning strategies are facing diverse challenges as data and

label inaccessibility increase. Training on few labeled data is a pivotal task and needs substantial resources. Uncovering the underlying graphical structure of EHRs brings a motivating perspective and informative prerequisites to analyzing patient data. As a downstream analysis method, GAN-boosted SSL uses a graphical structure and greatly improves learning quality in label-deficient situations. GANs with refined loss also meet the demands of deidentification and decent data fidelity under multiple-source data-sharing circumstances. This combination achieves impressive performance on prediction metrics, data quality, and protection from compromising attackers over various data sets, while popular machine learning methods encounter obstacles to sufficient training. This study indicates the potential of discovering the structural features that underlie the data instead of merely feeding models coordinated data sets and using unlabeled data when label deficiency occurs.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (82172069), the Major Scientific Project of Zhejiang Lab (2020ND8AD01), and the Fundamental Research Funds for the Central Universities (226-2023-00050). We owe thanks to the staff of the National Cancer Institute and each member involved in the Surveillance, Epidemiology and End Results program.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Graph structure definition and semisupervised learning on graph formularization.
[DOCX File , 82 KB - medinform_v11i1e47862_app1.docx]

References

1. Hripcsak G, Duke J, Shah N, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. Stud Health Technol Inform 2015;216:574-578 [FREE Full text] [doi: 10.3233/978-1-61499-564-7-574] [Medline: 26262116]
2. Banda JM, Seneviratne M, Hernandez-Boussard T, Shah NH. Advances in electronic phenotyping: From rule-based definitions to machine learning models. Annu Rev Biomed Data Sci 2018 Jul;1(1):53-68 [FREE Full text] [doi: 10.1146/annurev-biodatasci-080917-013315] [Medline: 31218278]
3. Xiao C, Choi E, Sun J. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. J Am Med Inform Assoc 2018 Oct 01;25(10):1419-1428 [FREE Full text] [doi: 10.1093/jamia/ocy068] [Medline: 29893864]
4. Blumenthal D, Tavenner M. The "meaningful use" regulation for electronic health records. N Engl J Med 2010 Aug 05;363(6):501-504. [doi: 10.1056/NEJMp1006114] [Medline: 20647183]
5. Gottesman O, Kuivaniemi H, Tromp G, Faucett WA, Li R, Manolio TA, eMERGE Network. The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. Genet Med 2013 Oct;15(10):761-771 [FREE Full text] [doi: 10.1038/gim.2013.72] [Medline: 23743551]
6. Summary of the HIPAA Privacy Rule. US Department of Health and Human Services. URL: https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html [accessed 2022-10-19]
7. El Emam K, Rodgers S, Malin B. Anonymising and sharing individual patient data. BMJ 2015 Mar 20;350:h1139 [FREE Full text] [doi: 10.1136/bmj.h1139] [Medline: 25794882]
8. Herbert RD, Kasza J, Bø K. Analysis of randomised trials with long-term follow-up. BMC Med Res Methodol 2018 May 29;18(1):48 [FREE Full text] [doi: 10.1186/s12874-018-0499-5] [Medline: 29843614]
9. Calzetta L, Matera MG, Goldstein MF, Fairweather WR, Howard WW, Cazzola M, et al. A long-term clinical trial on the efficacy and safety profile of doxofylline in Asthma: The LESDA study. Pulm Pharmacol Ther 2020 Feb;60:101883 [FREE Full text] [doi: 10.1016/j.pupt.2019.101883] [Medline: 31884206]
10. Welinder PP. Online crowdsourcing: rating annotators and obtaining cost-effective labels. 2010 Presented at: IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops; June 13-18, 2010; San Francisco, CA p. 25-32. [doi: 10.1109/cvprw.2010.5543189]
11. McLachlan S, Dube K, Gallagher T. Using the CareMap with health incidents statistics for generating the realistic synthetic electronic healthcare record. 2016 Presented at: IEEE International Conference on Healthcare Informatics (ICHI); October 4-7, 2016; Chicago, IL p. 439-448. [doi: 10.1109/ichi.2016.83]
12. Walonoski J, Kramer M, Nichols J, Quina A, Moesel C, Hall D, et al. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. J Am Med Inform Assoc 2018 Mar 01;25(3):230-238 [FREE Full text] [doi: 10.1093/jamia/ocx079] [Medline: 29025144]

13. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial networks. Commun. ACM 2020 Oct 22;63(11):139-144 [FREE Full text] [doi: 10.1145/3422622]
14. Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks. ArXiv. Preprint posted online Nov 19, 2015 2023. [doi: 10.48550/arXiv.1511.06434]
15. Choi E, Biswal S, Malin B. Generating multi-label discrete patient records using generative adversarial networks. 2017 Presented at: The 2nd Machine Learning for Healthcare Conference; August 18-19, 2017; Boston, MA p. 2017-2305 URL: https://proceedings.mlr.press/v68/choi17a.html
16. Baowaly M, Lin C, Liu C, Chen K. Synthesizing electronic health records using improved generative adversarial networks. J Am Med Inform Assoc 2019 Mar 01;26(3):228-241 [FREE Full text] [doi: 10.1093/jamia/ocy142] [Medline: 30535151]
17. Xie L, Wang S, Wang F, Zhou J. Differentially private generative adversarial network. ArXiv. Preprint posted online Feb 19, 2018 2023. [doi: 10.48550/arXiv.1802.06739]
18. Zhang X, Ji S, Wang T. Differentially private releasing via deep generative model (technical report). ArXiv. Preprint posted online Jan 5, 2018 2023 (forthcoming) [FREE Full text] [doi: 10.48550/arXiv.1801.01594]
19. Beaulieu-Jones BK, Greene CS, Pooled Resource Open-Access ALS Clinical Trials Consortium. Semi-supervised learning of the electronic health record for phenotype stratification. J Biomed Inform 2016 Dec;64:168-178 [FREE Full text] [doi: 10.1016/j.jbi.2016.10.007] [Medline: 27744022]
20. Estiri H, Murphy SN. Semi-supervised encoding for outlier detection in clinical observation data. Comput Methods Programs Biomed 2019 Nov;181:104830 [FREE Full text] [doi: 10.1016/j.cmpb.2019.01.002] [Medline: 30658851]
21. Che Z, Cheng Y, Zhai S, Sun Z, Liu Y. Boosting deep learning risk prediction with generative adversarial networks for electronic health records. 2017 Presented at: IEEE International Conference on Data Mining (ICDM); November 18-21, 2017; New Orleans, LA p. 787-792. [doi: 10.1109/icdm.2017.93]
22. van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res 2008;9(86):2579-2605 [FREE Full text]
23. van der Maaten L. Accelerating t-SNE using tree-based algorithms. J Mach Learn Res 2014;15(1):3221-3245 [FREE Full text]
24. Khan A, Uddin S, Srinivasan U. Comorbidity network for chronic disease: A novel approach to understand type 2 diabetes progression. Int J Med Inform 2018 Jul;115:1-9. [doi: 10.1016/j.ijmedinf.2018.04.001] [Medline: 29779710]
25. Defferrard M, Bresson X, Vandergheynst P. Convolutional neural networks on graphs with fast localized spectral filtering. 2016 Presented at: 30th Conference on Neural Information Processing Systems (NIPS 2016); December 5-10, 2016; Barcelona, Spain URL: https://proceedings.neurips.cc/paper/2016/hash/04df4d434d481c5bb723be1b6df1ee65-Abstract.html [doi: 10.48550/arXiv.1606.09375]
26. Hamilton W, Ying R, Leskovec J. Inductive representation learning on large graphs. 2017 Dec Presented at: 31st Conference on Neural Information Processing Systems (NIPS); December 4-9, 2017; Long Beach, CA. [doi: 10.48550/arXiv.1706.02216]
27. Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. 2017 Presented at: International Conference on Learning Representations; April 24-26, 2017; Toulon, France. [doi: 10.48550/arXiv.1609.02907]
28. Veličković P, Cucurull G, Casanova A, Romero-Soriano A, Liò P, Bengio Y. Graph attention networks. 2018 Presented at: International Conference on Learning Representations; Apr 30-May 3, 2018; Vancouver, BC URL: https://arxiv.org/abs/1710.10903
29. De Cao N, Kipf T. Molgan: An implicit generative model for small molecular graphs. ArXiv. Preprint posted online May 30, 2018 2018 May:1-13 (forthcoming). [doi: 10.48550/arXiv.1805.11973]
30. Bojchevski AOS, Zügner D, Günnemann S. Netgan: Generating graphs via random walks. 2018 Presented at: The 35th International Conference on Machine Learning; July 10-15, 2018; Stockholm, Sweden p. 610-619. [doi: 10.48550/arXiv.1803.00816]
31. Ding M, Tang J, Zhang J. Semi-supervised learning on graphs with generative adversarial nets. 2018 Presented at: CIKM '18: Proceedings of the 27th ACM International Conference on Information and Knowledge Management; October 22-26, 2018; Torino, Italy p. 913-922. [doi: 10.1145/3269206.3271768]
32. Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks. 2017 Presented at: International Conference on Machine Learning; August 6-11, 2017; Toulon, France p. 214-223 URL: https://proceedings.mlr.press/v70/arjovsky17a.html
33. Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville A. Improved training of Wasserstein GANs. 2017 Presented at: The 31st International Conference on Neural Information Processing Systems; December 4-9, 2017; Long Beach, CA p. 5769-5779 URL: https://proceedings.neurips.cc/paper_files/paper/2017/hash/892c3b1c6dccd52936e27cbd0ff683d6-Abstract.html
34. Wang H, Wang J, Wang J, Zhao M, Zhang W, Zhang F, et al. GraphGAN: graph representation learning with generative adversarial nets. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2018 Presented at: Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18); February 2–7, 2018; New Orleans, LA. [doi: 10.1609/aaai.v32i1.11872]
35. Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks. 2017 Presented at: 34th International Conference on Machine Learning; August 6-11, 2017; Sydney, Australia p. 214-223 URL: https://proceedings.mlr.press/v70/arjovsky17a.html

36. De Cao DZ, Kipf T. MolGAN: An implicit generative model for small molecular graphs. ArXiv. Preprint posted online May 30, 2018 2018:1-13. [doi: 10.48550/arXiv.1805.11973]
37. Bojchevski A, Shchur O, Zügner D, Günnemann S. Netgan: Generating graphs via random walks. 2018 Presented at: Proceedings of the 35th International Conference on Machine Learning; July 10-15, 2018; Stockholm, Sweden p. 610-619 URL: https://proceedings.mlr.press/v80/bojchevski18a.html
38. Tang J, Qu M, Wang M, Zhang M, Yan J, Mei Q. LINE: Large-scale Information Network Embedding. 2015 Presented at: Proceedings of the 24th International Conference on World Wide Web; May 18-22, 2015; Florence, Italy p. 1067. [doi: 10.1145/2736277.2741093]
39. Grandvalet Y, Bengio Y. Semi-supervised learning by entropy minimization. In: Advances in Neural Information Processing Systems. 2004 Presented at: The 17th International Conference on Neural Information Processing Systems; December 1, 2004; Vancouver, BC p. 529-536 URL: https://proceedings.neurips.cc/paper/2004/hash/96f2b50b5d3613adf9c27049b2a888c7-Abstract.html
40. Springenberg JT. Unsupervised and semi-supervised learning with categorical generative adversarial networks. 2016 Presented at: International Conference on Learning Representations; May 2-4, 2016; San Juan, Puerto Rico. [doi: 10.48550/arXiv.1511.06390]
41. Eby E, Hardwick C, Yu M, Gelwicks S, Deschamps K, Xie J, et al. Predictors of 30 day hospital readmission in patients with type 2 diabetes: a retrospective, case-control, database study. Curr Med Res Opin 2015 Jan;31(1):107-114. [doi: 10.1185/03007995.2014.981632] [Medline: 25369567]
42. Secondary Surveillance, Epidemiology, and End Results. National Cancer Institute. URL: https://www.seer.cancer.gov [accessed 2023-05-16]
43. Li L, Cheng W, Glicksberg BS, Gottesman O, Tamler R, Chen R, et al. Identification of type 2 diabetes subgroups through topological analysis of patient similarity. Sci Transl Med 2015 Oct 28;7(311):311ra174 [FREE Full text] [doi: 10.1126/scitranslmed.aaa9364] [Medline: 26511511]
44. Zhang J, Feng W, Yuan T, Wang J, Sangaiah AK. SCSTCF: spatial-channel selection and temporal regularized correlation filters for visual tracking. Appl Soft Comput 2022 Mar;118:108485. [doi: 10.1016/j.asoc.2022.108485]
45. Zhang J, Sun J, Wang J, Li Z, Chen X. An object tracking framework with recapture based on correlation filters and Siamese networks. Comput Electr Eng 2022 Mar;98:107730. [doi: 10.1016/j.compeleceng.2022.107730]
46. Hwang U, Jung D, Yoon S. HexaGAN: generative adversarial nets for real world classification. 2019 Presented at: ICML 2019: 36th International Conference on Machine Learning; June 10-15, 2019; Long Beach, CA. [doi: 10.1007/978-3-658-40442-0_9]

Abbreviations

AUC: area under the receiver operating characteristics curve
DWPre: dimensionwise precision
DWPro: dimensionwise probability
EHR: electronic health record
eMERGE: Electronic Medical Records and Genomics
GAN: generative adversarial network
GCN: graph convolutional network
GPU: graphics processing unit
GSSL: graph-base semisupervised learning
GSSL: graph-based semisupervised learning
k-NN: k-nearest neighbors
LP: label propagation
LR: logistic regression
RF: random forest
RS-EHR: realistic synthesized electronic health record
SAHZU-CRC: Second Affiliated Hospital of Zhejiang University Colorectal Cancer
SEER-CRC: Surveillance, Epidemiology, and End Results Colorectal Cancer
SEER-OVC: Surveillance, Epidemiology, and End Results Ovarian Cancer
SL: supervised learning
SSL: semisupervised learning
SVM: support vector machine
UCI-T2D: University of California Irvine Machine Learning Repository Type 2 Diabetes 30-Day Readmission

Edited by G Eysenbach, C Lovis; submitted 04.04.23; peer-reviewed by M Gong, W Zhu; comments to author 03.05.23; revised version received 11.05.23; accepted 12.05.23; published 13.06.23.

Please cite as:
Li R, Tian Y, Shen Z, Li J, Li J, Ding K, Li J
Improving an Electronic Health Record–Based Clinical Prediction Model Under Label Deficiency: Network-Based Generative Adversarial Semisupervised Approach
JMIR Med Inform 2023;11:e47862
URL: https://medinform.jmir.org/2023/1/e47862
doi:10.2196/47862
PMID:37310778

Original Paper

A Large Language Model Screening Tool to Target Patients for Best Practice Alerts: Development and Validation

Thomas Savage[1], MD; John Wang[2], MD; Lisa Shieh[1], MD, PhD

[1]Division of Hospital Medicine, Department of Medicine, Stanford University, Palo Alto, CA, United States

[2]Divison of Gastroenterology and Hepatology, Department of Medicine, Stanford University, Palo Alto, CA, United States

Corresponding Author:
Thomas Savage, MD
Division of Hospital Medicine
Department of Medicine
Stanford University
300 Pasteur Drive
Palo Alto, CA, 94304
United States
Phone: 1 6507234000
Email: tsavage@stanford.edu

Abstract

Background: Best Practice Alerts (BPAs) are alert messages to physicians in the electronic health record that are used to encourage appropriate use of health care resources. While these alerts are helpful in both improving care and reducing costs, BPAs are often broadly applied nonselectively across entire patient populations. The development of large language models (LLMs) provides an opportunity to selectively identify patients for BPAs.

Objective: In this paper, we present an example case where an LLM screening tool is used to select patients appropriate for a BPA encouraging the prescription of deep vein thrombosis (DVT) anticoagulation prophylaxis. The artificial intelligence (AI) screening tool was developed to identify patients experiencing acute bleeding and exclude them from receiving a DVT prophylaxis BPA.

Methods: Our AI screening tool used a BioMed-RoBERTa (Robustly Optimized Bidirectional Encoder Representations from Transformers Pretraining Approach; AllenAI) model to perform classification of physician notes, identifying patients without active bleeding and thus appropriate for a thromboembolism prophylaxis BPA. The BioMed-RoBERTa model was fine-tuned using 500 history and physical notes of patients from the MIMIC-III (Medical Information Mart for Intensive Care) database who were not prescribed anticoagulation. A development set of 300 MIMIC patient notes was used to determine the model's hyperparameters, and a separate test set of 300 patient notes was used to evaluate the screening tool.

Results: Our MIMIC-III test set population of 300 patients included 72 patients with bleeding (ie, were not appropriate for a DVT prophylaxis BPA) and 228 without bleeding who were appropriate for a DVT prophylaxis BPA. The AI screening tool achieved impressive accuracy with a precision-recall area under the curve of 0.82 (95% CI 0.75-0.89) and a receiver operator curve area under the curve of 0.89 (95% CI 0.84-0.94). The screening tool reduced the number of patients who would trigger an alert by 20% (240 instead of 300 alerts) and increased alert applicability by 14.8% (218 [90.8%] positive alerts from 240 total alerts instead of 228 [76%] positive alerts from 300 total alerts), compared to nonselectively sending alerts for all patients.

Conclusions: These results show a proof of concept on how language models can be used as a screening tool for BPAs. We provide an example AI screening tool that uses a HIPAA (Health Insurance Portability and Accountability Act)–compliant BioMed-RoBERTa model deployed with minimal computing power. Larger models (eg, Generative Pre-trained Transformers–3, Generative Pre-trained Transformers–4, and Pathways Language Model) will exhibit superior performance but require data use agreements to be HIPAA compliant. We anticipate LLMs to revolutionize quality improvement in hospital medicine.

(JMIR Med Inform 2023;11:e49886) doi:10.2196/49886

KEYWORDS

large language models; language models; language model; EHR; health record; health records; quality improvement; Artificial Intelligence; Natural Language Processing

Introduction

Large language models (LLMs) are an exciting development in the field of natural language processing and present tremendous potential for application to clinical medicine. Language models such as Bidirectional Encoder Representations from Transformers (BERT), Robustly Optimized BERT Pretraining Approach (RoBERTa), and Generative Pre-trained Transformers (GPT) can perform complex tasks such as text classification, question answering, and text generation, which have the potential to augment physicians in providing care for patients. A clinical application of significant potential is LLM optimization of quality improvement electronic health record (EHR) Best Practice Alerts (BPAs).

EHR BPAs are powerful tools to improve patient care outcomes. Their widespread use has demonstrated better adherence to clinical guidelines, fewer medication errors, improved diabetes management, more appropriate antimicrobial prescribing, and higher rates of ambulatory preventive care, among others [1]. Nevertheless, despite these benefits, the overuse of BPAs can cause alarm fatigue and desensitization that can even lead to patient harm, as described by the Joint Commission Alert System Safety Report of 2014 [2,3]. Current BPAs are often overconservative, broadly applied to all possible patients, resulting in up to 49% to 96% of alerts being overridden or ignored [4]. A BPA will frequently be obviously inappropriate or not applicable to the patient, resulting in the physician becoming desensitized to the alert and more likely to ignore a valid alert in the future [5,6]. LLMs offer an opportunity to screen appropriate patients for BPAs.

In this paper, we propose how language models can be leveraged to read a physician's note and screen whether a patient is appropriate for a BPA. We examine an example case using a BioMed-RoBERTa artificial intelligence (AI) screening tool that selectively identifies patients who are appropriate for a deep vein thrombosis prophylaxis. Our screening tool reads each history and physical note to identify whether a patient has active bleeding (a contraindication to anticoagulation) and excludes those patients with bleeding from receiving a deep vein thrombosis prophylaxis BPA. We hope our methods will encourage the use of language models to screen patients for BPAs and improve EHR workflow.

Methods

Patient Note Data Set

The MIMIC-III (Medical Information Mart for Intensive Care) data set is publicly available and consists of more than 60,000 intensive care unit admissions from the Brigham and Women's Hospital system [7-9]. The data set includes EHR-equivalent patient data, including physician notes and medication administration data.

For this study, we selected history and physical notes from MIMIC-III for patients who did not receive an anticoagulant (therapeutic or prophylactic) at the time of admission. We selected the first 1100 history and physical notes in chronological time stamp order for inclusion in this study. Each note described a unique patient, meaning no 2 notes had the same patient identifier (subject_id). The subject_id and time stamp for each note included in our study were recorded and can be shared upon request with proper registry for the MIMIC database on the PhysioNet platform. Due to LLM token limits, patient notes were truncated to 2000 characters (RoBERTa models can receive a maximum of 512 tokens, which translates to roughly 2000 characters).

The MIMIC-III notes were then split into a training set (first 500 notes), development set (middle 300 notes), and test set (final 300 notes). The training set was used to fine-tune the AI model, the development set was used to determine model hyperparameters (learning rate, batch number, training epochs, and seed number), and the test set was used for screening tool evaluation. Cohort information can be shared upon request with registry for the MIMIC database on the PhysioNet platform.

In total, 2 physicians (TS and JW) reviewed all notes and labeled each note as either describing a patient with active bleeding or without active bleeding. Active bleeding was defined as the loss of blood from the vessels of the body either by documented visual or imaging evidence of bleeding as well as suspected bleeding documented by the physician author. These labels were used as the gold standard labels for model evaluation. If there was disagreement between labels assigned by the 2 reviewers, the case was discussed to reach a final label designation.

Language Model

The language model used in this screening tool was BioMed-RoBERTa [10] by AllenAI. BioMed-RoBERTa is a bidirectional transformer encoder based on the RoBERTa-base model published by Liu et al [11]. BioMed-RoBERTa continued pretraining beyond RoBERTa-base using 2.68 million scientific papers in the domains of biology and medicine from the Semantic Scholar corpus [10]. Therefore BioMed-RoBERTa has increased proficiency in the subjects of biology and medicine compared to RoBERTa-base.

The optimal hyperparameter settings were identified as 9 training epochs, a training batch size of 2, a learning rate of 4×10^{-5}, and a starting seed of 9. The full code can be referenced in Multimedia Appendix 1

The model performed classification for each history and physical note, classifying the note as either describing active bleeding or no active bleeding.

Code and Computing Environment

Model training and evaluation were completed in a PyCharm notebook using an Apple M2 GPU. Full code can be referenced in Multimedia Appendix 1.

Statistical Methods

Model classification performance was evaluated on a test set of 300 MIMIC patient notes. Classification performance was evaluated by a precision-recall area under the curve (AUC), a receiver operating characteristic (ROC) AUC, sensitivity, and specificity. Error was calculated using a replacement bootstraps method with 4000 bootstrapped populations from the test set. The full code can be referenced in Multimedia Appendix 1.

The statistic "increase in alert applicability" was calculated from equation 1 below.

Ethical Considerations

The MIMIC patient records database is available as an open repository for credential users registered within the PhysioNet platform. The patient records within the MIMIC database have been deidentified, and patient identifiers have been removed according to the HIPAA (Health Insurance Portability and Accountability Act) Safe Harbor provision. This study was approved and authorized by the PhysioNet Team.

The collection of patient information and creation of the MIMIC research resource was reviewed by the institutional review board at the Beth Israel Deaconess Medical Center (protocol ID 73104), who granted a waiver of informed consent and approved the data sharing initiative. Due to the waiver of informed consent and anonymous nature of the MIMIC data set, patients were not compensated for inclusion in the database or this research.

Results

The test set of 300 MIMIC patient notes included 72 patients experiencing acute bleeding and 228 patients not experiencing acute bleeding. The AI screening tool was able to achieve a precision-recall curve AUC of 0.82 (95% CI 0.75-0.89; Figure 1) and a ROC AUC of 0.89 (95% CI 0.84-0.94; Figure 2). Sensitivity was found to be 67% (95% CI 55%-77%) and specificity was found to be 95% (95% CI 92%-97%). Sensitivity, specificity, and confusion matrix data can be found in Table 1.

If this model were used to identify patients for a thromboembolism prophylaxis BPA in place of a nonselective strategy deploying an alert for all patients, this model would reduce the number of BPA alerts by 20% (240 alerts sent instead of 300) and increase applicability of the BPA by 14.8% (equation 1). This model misclassified 12 patients who did not have bleeding and would be appropriate to receive a thromboembolism prophylaxis BPA. This model captured 94.7% (216/228) of the population who would be appropriate for an alert. The model achieved a positive predictive value of 80% and a negative predictive value of 90%. Full results can be found in Table 1.

A review of notes of patients who were incorrectly classified found trends where the tool underperformed. Of the 24 patients misclassified as without bleeding, the model had difficulty identifying bleeding if it was a secondary complaint (6 patients), meaning if the patient primarily presented for another chief complaint and bleeding was incidentally noted. The model also had difficulty interpreting atypical or rare abbreviations that denoted bleeding (6 patients). This included recognizing "subdural" as an abbreviation for subdural hemorrhage (2 patients), "EBL" as an abbreviation for estimated blood loss (2 patients), and "SAH" as an abbreviation for subarachnoid hemorrhage (2 patients). For the 12 patients misclassified as with bleeding, the model had difficulty with negated bleeding phrases (7 patients), for example "denies melena," as well as notes describing chronic anemia (2 patients) or previous bleeding events in the distant past that were not currently active (2 patients).

Figure 1. Precision-recall curve for the BioMed-RoBERTa model. AUC: area under the curve; RoBERTa: Robustly Optimized Bidirectional Encoder Representations from Transformers Pretraining Approach.

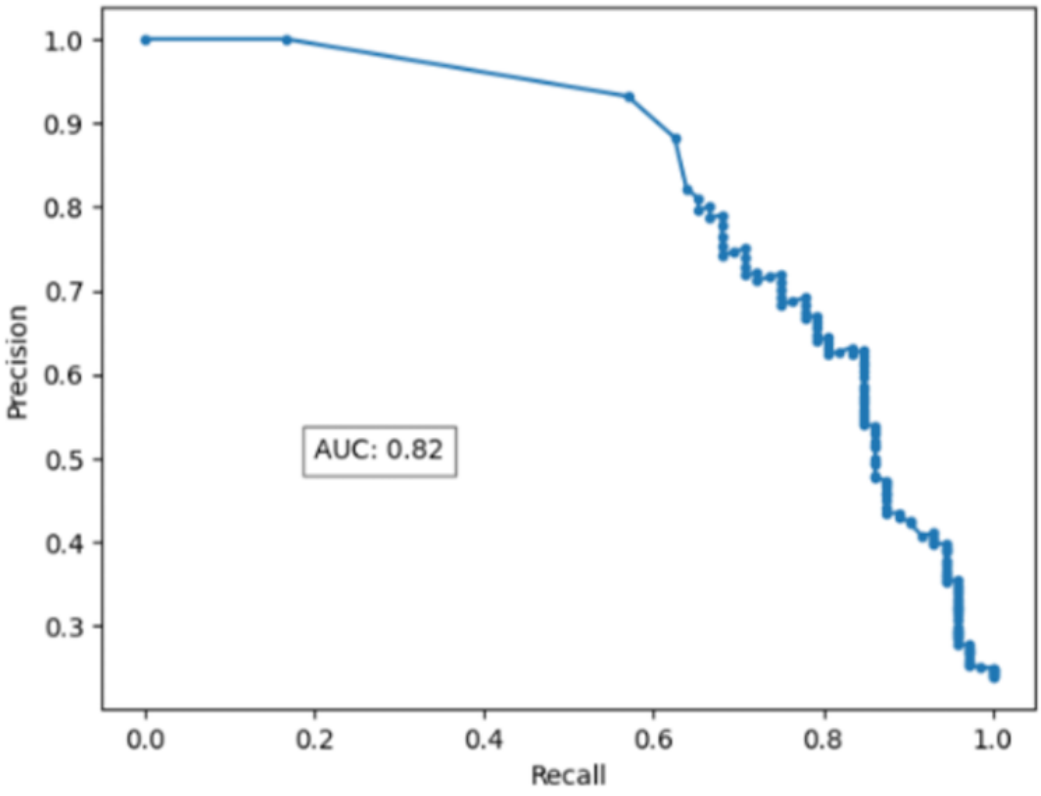


Figure 2. ROC curve for the BioMed-RoBERTa model. AUC: area under the curve; ROC: receiver operating characteristic; RoBERTa: Robustly Optimized Bidirectional Encoder Representations from Transformers Pretraining Approach.

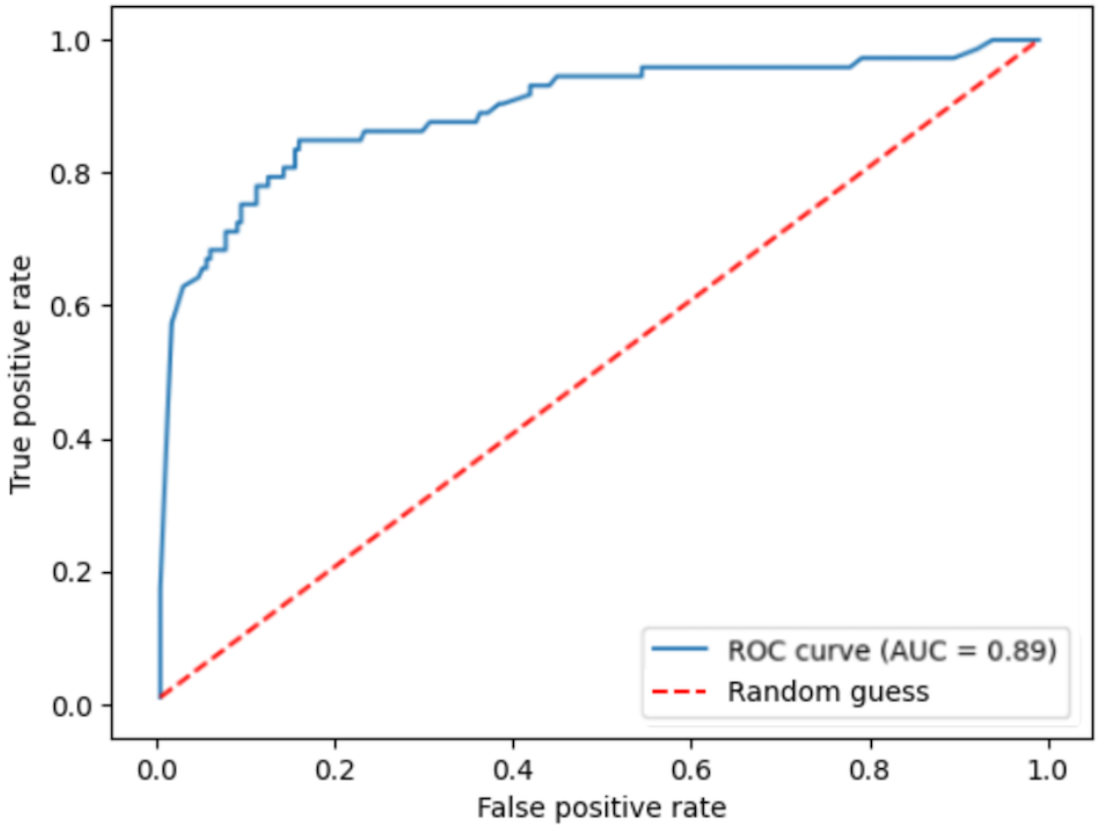


Table 1. Confusion matrix for the BioMed-RoBERTa[a] model's classification results.

LLM[b] classification	Physician label		
	Positive	Negative	Total
Positive	48	12	60
Negative	24	216	240
Total	72	228	300

[a]RoBERTa: Robustly Optimized Bidirectional Encoder Representations from Transformers Pretraining Approach.

[b]LLM: large language model.

Discussion

Principal Findings

This prototype AI screening tool shows how language models can be leveraged to optimize EHR BPAs. Our example screening tool reduced the number of patients who would trigger an alert by 20% while also increasing applicability of alerts by 14.8% compared to nonselectively sending alerts for all patients. The tool achieved impressive accuracy with a precision-recall AUC of 0.82 and ROC AUC of 0.89. We demonstrate how the ability for language models to interpret physician notes to classify patients offers a new method of targeting the most appropriate patients for a BPA with high accuracy.

Language model screening tools to target patients for BPAs are most appropriate when the alert's goal is to catch most target patients but the consequence of missing a few patients is not significant. As demonstrated by our example, language models are accurate but do occasionally misclassify patients. Example use cases appropriate for an AI screening tool would be identifying patients for discontinuation of telemetry monitoring, ascribing correct level of nursing care (inpatient vs observation), and encouraging best prescribing practices for blood products. Inappropriate examples would be situations such as medication interactions or isolation precautions. In those cases, a nonselective blanket rule-based BPA would be more appropriate.

The limitations of the screening tool evaluated in this study were its relatively small training set of 500 patient notes as well as the use of the BioMed-RoBERTa model rather than a larger model such as GPT-3 or GPT-4. Increasing the size of our training set would likely decrease many of the misclassification errors seen by our screening tool. Specifically, the misclassification errors due to misinterpretation of atypical or rare abbreviations for bleeding would be reduced with an increased training set size. A larger base model would also reduce many of the errors observed by our screening tool. RoBERTa models contain 110 million parameters trained on

160 GB of text data [11]. Larger models, such as GPT-3, consist of 175 billion parameters trained on over 40 TB of text data, nearly 10 times the size of RoBERTa, and demonstrate superior performance in negation detection and semantic understanding [12,13]. Therefore, a larger model would reduce the misclassification errors caused by misinterpretation of negated bleeding or previous bleeding events in the distant past. Our study chose BioMed-RoBERTa because its smaller size allows it to be trained on a local computing environment compliant with the HIPAA 1996 data privacy standards. Future investigations will need to secure the necessary data use agreements to use larger models (eg, GPT-3, GPT-4, or Pathways Language Model) with medical grade data, where we anticipate screening tool performance will be significantly improved.

Conclusions

In this paper, we proposed a new application for LLMs in medicine. Quality improvement BPAs can leverage LLMs to read physician notes and better identify patients for BPA alerts. We provide an example case which demonstrates the ability of a BioMed-RoBERTa to achieve impressive classification accuracy. We anticipate that as the field of clinical natural language processing continues to grow, with increasing access to larger language models, LLMs will revolutionize the field of clinical quality improvement.

Data Availability

The Python code used for this investigation is available in Multimedia Appendix 1. MIMIC (Medical Information Mart for Intensive Care) patient data is an open repository requiring registration with the PhysioNet platform. Cohort information of the history and physical notes used to train, develop, and test our screening tool in can be shared upon request with proper registry for the MIMIC database on the PhysioNet platform.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Python code for training and evaluating the large language model (Robustly Optimized Bidirectional Encoder Representations from Transformers Pretraining Approach; RoBERTa) screening tool used in this investigation.
[PDF File (Adobe PDF File), 46 KB - medinform_v11i1e49886_app1.pdf]

References

1. Ancker JS, Kern LM, Edwards A, Nosal S, Stein DM, Hauser D, et al. Associations between healthcare quality and use of electronic health record functions in ambulatory care. J Am Med Inform Assoc 2015;22(4):864-871 [FREE Full text] [doi: 10.1093/jamia/ocv030] [Medline: 25896648]
2. Ancker JS, Edwards A, Nosal S, Hauser D, Mauer E, Kaushal R, et al. Effects of workload, work complexity, and repeated alerts on alert fatigue in a clinical decision support system. BMC Med Inform Decis Mak 2017;17(1):36 [FREE Full text] [doi: 10.1186/s12911-017-0430-8] [Medline: 28395667]
3. Mitka M. Joint commission warns of alarm fatigue: multitude of alarms from monitoring devices problematic. JAMA 2013;309(22):2315-2316. [doi: 10.1001/jama.2013.6032] [Medline: 23757063]
4. van der Sijs H, Aarts J, Vulto A, Berg M. Overriding of drug safety alerts in computerized physician order entry. J Am Med Inform Assoc 2006;13(2):138-147 [FREE Full text] [doi: 10.1197/jamia.M1809] [Medline: 16357358]
5. Jaspers T, van Essenberg MD, Maat B, Durian M, van den Berg R, van den Bemt P. A multifaceted clinical decision support intervention to improve adherence to thromboprophylaxis guidelines. Int J Clin Pharm 2021;43(5):1327-1336 [FREE Full text] [doi: 10.1007/s11096-021-01254-x] [Medline: 33709383]
6. Spirk D, Stuck AK, Hager A, Engelberger RP, Aujesky D, Kucher N. Electronic alert system for improving appropriate thromboprophylaxis in hospitalized medical patients: a randomized controlled trial. J Thromb Haemost 2017;15(11):2138-2146 [FREE Full text] [doi: 10.1111/jth.13812] [Medline: 28836340]
7. Johnson A, Pollard T, Mark R. MIMIC-III clinical database. PhysioNet. 2015. URL: https://doi.org/10.13026/C2XW26 [accessed 2023-11-07]
8. Johnson AEW, Pollard TJ, Shen L, Lehman LWH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016;3:160035 [FREE Full text] [doi: 10.1038/sdata.2016.35] [Medline: 27219127]
9. Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation 2000;101(23):E215-E220 [FREE Full text] [doi: 10.1161/01.cir.101.23.e215] [Medline: 10851218]
10. allenai/biomed_roberta_base. Hugging Face. 2023. URL: https://huggingface.co/allenai/biomed_roberta_base [accessed 2023-07-08]
11. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: a robustly optimized BERT pretraining approach. arXiv :1-13 Preprint posted online on July 26, 2019. [FREE Full text] [doi: 10.48550/arXiv.1907.11692]

12. Nguyen HT, Goebel R, Toni F, Stathis K, Satoh K. A negation detection assessment of GPTs: analysis with the xNot360 dataset. arXiv :1-8 Preprint posted online on June 29, 2023 [FREE Full text] [doi: 10.48550/arXiv.2306.16638]
13. Tejada GNC, Scholtes J, Spanakis G. A study of BERT's processing of negations to determine sentiment. 2021 Presented at: 33rd Benelux Conference on Artificial Intelligence and 30th Belgian-Dutch Conference on Machine Learning; November 10-12, 2021; Belval, Esch-sur-Alzette, Luxembourg p. 47-59 URL: https://dke.maastrichtuniversity.nl/jerry.spanakis/wp-content/uploads/2021/11/NegationBERT.pdf

Abbreviations

AI: artificial intelligence
AUC: area under the curve
BERT: Bidirectional Encoder Representations from Transformers
BPA: Best Practice Alert
DVT: deep vein thrombosis
EHR: electronic health record
GPT: Generative Pre-trained Transformers
HIPAA: Health Insurance Portability and Accountability Act
LLM: large language model
MIMIC: Medical Information Mart for Intensive Care
RoBERTa: Robustly Optimized Bidirectional Encoder Representations from Transformers Pretraining Approach
ROC: receiver operating characteristic

Edited by C Lovis; submitted 12.06.23; peer-reviewed by R Marshall, J Zeng, J Kors; comments to author 06.07.23; revised version received 04.08.23; accepted 24.10.23; published 27.11.23.

Please cite as:
Savage T, Wang J, Shieh L
A Large Language Model Screening Tool to Target Patients for Best Practice Alerts: Development and Validation
JMIR Med Inform 2023;11:e49886
URL: https://medinform.jmir.org/2023/1/e49886
doi:10.2196/49886
PMID:38010803

Original Paper

Applications of the Natural Language Processing Tool ChatGPT in Clinical Practice: Comparative Study and Augmented Systematic Review

Nikolas Schopow[1], MBA, Dr med; Georg Osterhoff[1], Dr med; David Baur[1], Dr med

Department for Orthopedics, Trauma Surgery and Plastic Surgery, University Hospital Leipzig, Leipzig, Germany

Corresponding Author:
Nikolas Schopow, MBA, Dr med
Department for Orthopedics, Trauma Surgery and Plastic Surgery
University Hospital Leipzig
Liebigstrasse 20
Leipzig, 04103
Germany
Phone: 49 341 97 ext 17300
Email: schopow@medizin.uni-leipzig.de

Abstract

Background: This research integrates a comparative analysis of the performance of human researchers and OpenAI's ChatGPT in systematic review tasks and describes an assessment of the application of natural language processing (NLP) models in clinical practice through a review of 5 studies.

Objective: This study aimed to evaluate the reliability between ChatGPT and human researchers in extracting key information from clinical articles, and to investigate the practical use of NLP in clinical settings as evidenced by selected studies.

Methods: The study design comprised a systematic review of clinical articles executed independently by human researchers and ChatGPT. The level of agreement between and within raters for parameter extraction was assessed using the Fleiss and Cohen κ statistics.

Results: The comparative analysis revealed a high degree of concordance between ChatGPT and human researchers for most parameters, with less agreement for study design, clinical task, and clinical implementation. The review identified 5 significant studies that demonstrated the diverse applications of NLP in clinical settings. These studies' findings highlight the potential of NLP to improve clinical efficiency and patient outcomes in various contexts, from enhancing allergy detection and classification to improving quality metrics in psychotherapy treatments for veterans with posttraumatic stress disorder.

Conclusions: Our findings underscore the potential of NLP models, including ChatGPT, in performing systematic reviews and other clinical tasks. Despite certain limitations, NLP models present a promising avenue for enhancing health care efficiency and accuracy. Future studies must focus on broadening the range of clinical applications and exploring the ethical considerations of implementing NLP applications in health care settings.

(JMIR Med Inform 2023;11:e48933) doi:10.2196/48933

KEYWORDS

natural language processing; clinical practice; systematic review; healthcare; health care; GPT-3; GPT-4; large language models; artificial intelligence; machine learning; clinical decision support systems; language model; NLP; ChatGPT; systematic; review methods; review methodology; text; unstructured; extract; extraction

Introduction

The following manuscript was augmented by ChatGPT (versions 3.5 and 4.0; OpenAI [1]). ChatGPT-generated text is shown in Roman (unitalicized) font and has not been altered. Any modifications to the generated text, including corrections to sources or information, are explicitly indicated. Any text added or revised by human authors is shown in italics. All in-text reference citations have been reformatted to adhere to the journal's style preferences.

Natural Language Processing (NLP) has emerged as a powerful tool in recent years, enabling the processing and analysis of vast amounts of unstructured textual data in various domains,

including healthcare and clinical practice [2] *(added [3])*. The application of NLP techniques in clinical settings has the potential to revolutionize the way medical professionals manage and analyze patient information, leading to improved patient outcomes, reduced costs, and increased efficiency in medical decision-making [4] *(added [5])*.

In clinical practice, NLP can facilitate various tasks, such as disease diagnosis, treatment decision support, automation of clinical tasks, and data mining [6]. For instance, NLP algorithms have been used to screen and identify patients at risk for specific conditions [7], aid in the diagnosis of diseases by analyzing electronic health records (EHRs) [original: Demner-Fushman, D., & Chapman, W. W. (2017)] *(new [8])*, provide decision support in treatment planning [original: Wang, Y et al. (2017)] *(new [9])*, and automate routine clinical tasks [original: Devlin, J. et al. (2019)] *(new [10])*. Furthermore, NLP has been employed in the analysis of large-scale medical literature to identify trends, generate hypotheses, and inform clinical decision-making [original: Brown, T. B. et al. (2020)] *(new [11])*.

Recent advancements in NLP, particularly the introduction of transformer-based models like Bidirectional Encoder Representations from Transformers (BERT) [original: Vaswani, A et al. (2017)] *(new [12])*, Generative Pre-trained Transformers (GPT) [original: Lee et al, 2020] *(new [13])*, and their variants *(moved [14])*, have significantly improved the performance of NLP tasks, including information extraction, question-answering, and text summarization. Transformer networks leverage attention mechanisms, allowing them to learn contextual relationships between words in a given text, thus enabling a more nuanced understanding of the input data [15].

Transformer-based models like BERT and GPT work using a self-attention mechanism, allowing them to focus on relevant words in a sentence, thus capturing contextual information efficiently [16]. This approach enables precise understanding of semantic relationships, making these models adept at tasks such as named entity recognition and text summarization [12].

However, these models have limitations. The attention mechanism is resource-intensive, potentially limiting their use in constrained environments [17]. Furthermore, while able to generate plausible-sounding outputs, they may occasionally produce nonsensical or incorrect results (called “artificial hallucination”), which emphasizes the need for careful interpretation [original: McCoy et al. 2019] *(new [18])*.

These models have been successfully applied to various healthcare-related tasks, including biomedical literature mining [19], clinical concept extraction [20] *(added [21])*, and predicting patient outcomes [original: Nye et al. 2018] *(new [22])*.

Large language models (LLM) represent the cutting-edge of NLP, demonstrating exceptional performance in various tasks by leveraging their extensive pre-training on vast textual data [original: Smith et al., 2022] *(new [23])*.

Yet, despite the notable advancements in NLP and LLMs, traditional systematic reviews continue to pose significant limitations [24] *(added [25])*. Traditional approaches to systematic reviews are often labor-intensive and time-consuming, involving manual screening of literature and information extraction [26]. Such processes are not only susceptible to human error [27] but also struggle to cope with the exponential increase in available medical literature [28]. The extensive and complex nature of medical data, combined with the ever-evolving landscape of clinical research, presents a substantial challenge to traditional systematic review methods [29] *(added [30])*. Thus, there is a pressing need for more sophisticated and automated solutions, such as those provided by NLP, to handle the growing volume and complexity of medical literature [31].

In light of these developments, we aim to conduct a systematic review aided by NLP, specifically leveraging the capabilities of transformer-based models like GPT, to synthesize the existing literature on the application of NLP in clinical practice. Our review will focus on studies published between January 2020 and the present, evaluating the performance, implementation, and impact of NLP techniques in clinical settings. By integrating NLP into the systematic review process, we aim to increase the efficiency and accuracy of the review, enabling the identification of relevant studies, extraction of key information, and synthesis of findings in a more streamlined manner [32].

LLMs have been gaining traction in both social media and the scientific community. We compared the results of human researchers *(with a research experience of >7 years)* versus ChatGPT (Versions 3.5 and 4.0) in an artificial intelligence augmented systematic review. The goal was to explore the usefulness and limitations of LLMs in clinical practice, medical research and writing publications.

The main aim was to evaluate how effectively and reliably ChatGPT could support the process of conducting a medical systematic review, while also identifying potential issues and offering insights into the rapidly evolving field of artificial intelligence.

Methods

Overview

The task of conducting a systematic review was augmented using ChatGPT. ChatGPT was used for general considerations in conducting a systematic review; determining MeSH (Medical Subject Headings) terms; title, abstract, and full-text screening; limited data extraction; and text generation.

This manuscript was generated in several sections; therefore, modifications for better readability—for example, the order of text sections, numbering of references, and the use of abbreviations—are not shown. Relevant conversations with ChatGPT are provided in Figure 1 and Multimedia Appendices 1-16.

Our systematic review followed the guidelines provided by the Cochrane Handbook for Systematic Reviews of Interventions [33] *(added [34])* and the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement [25]. *The PRISMA flowchart is shown in Figure 2 and the PRISMA checklist in Multimedia Appendix 17.*

Figure 1. An exemplary prompt and response from ChatGPT as a multistep answer for prompt generation for the abstract text module.

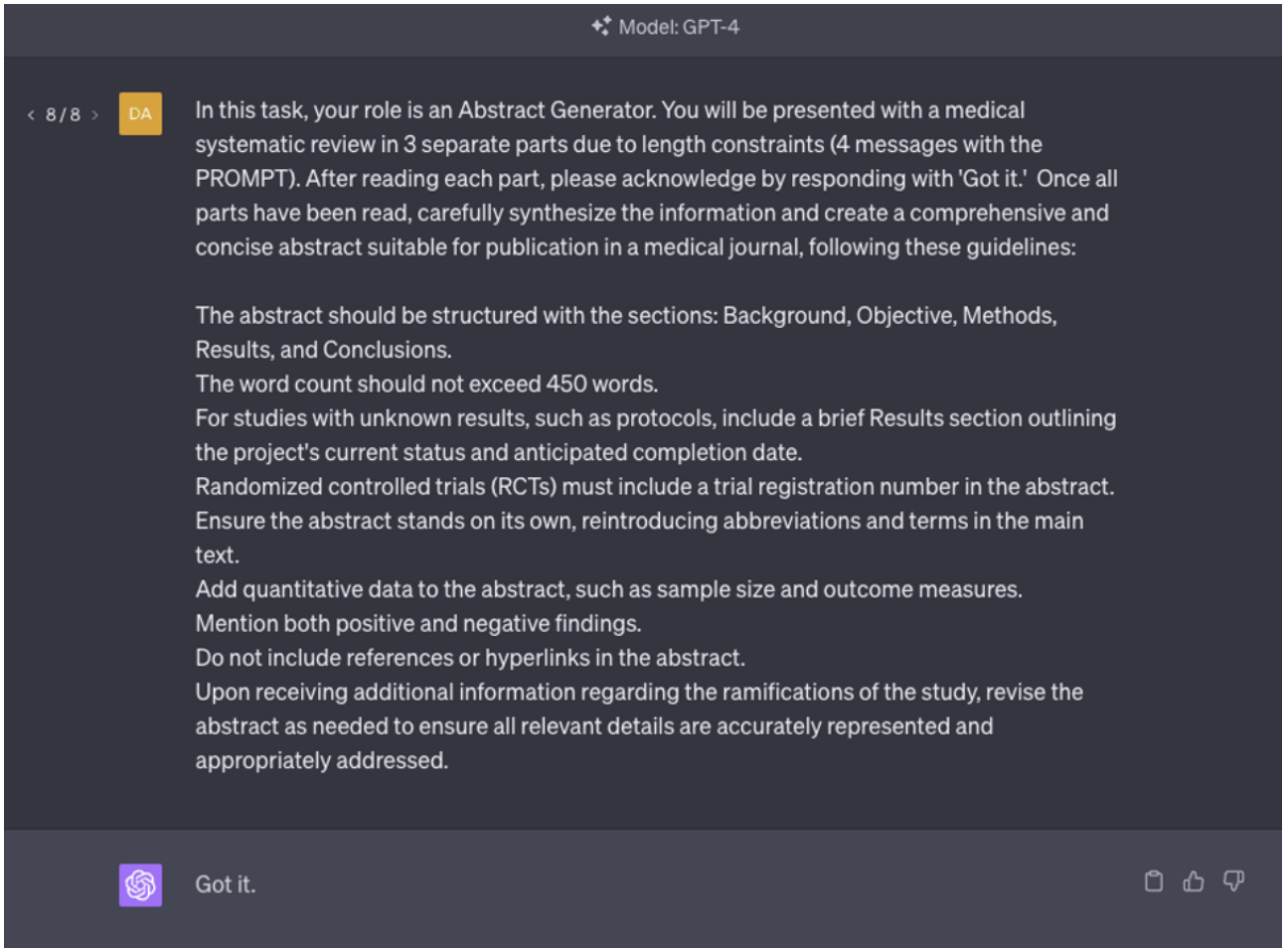
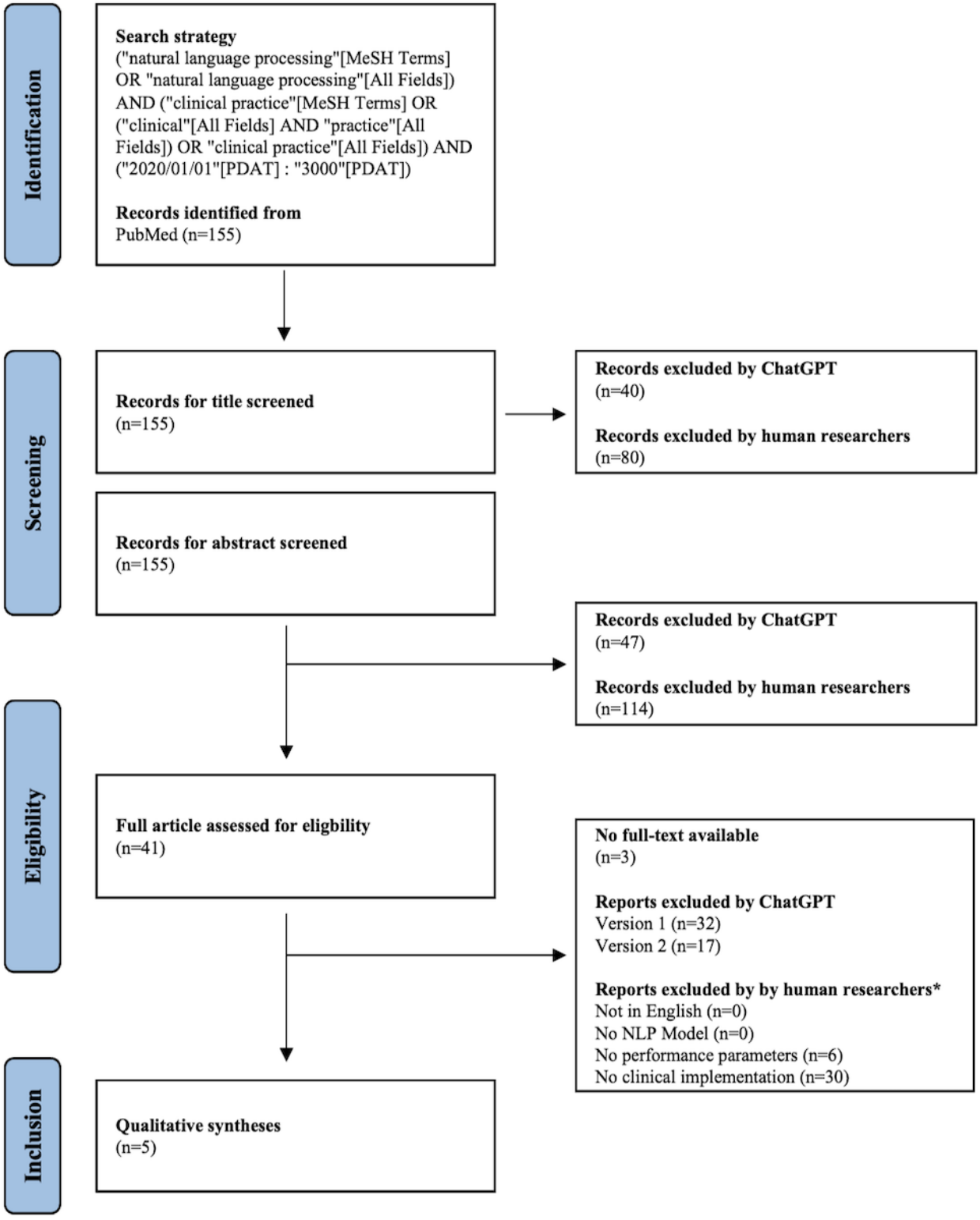


Figure 2. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flowchart. *In consensus, multiple reasons possible. MeSH: Medial Subject Headings; NLP: natural language processing.



Search Strategy

We utilized ChatGPT 3.5 legacy (Version Jan 30, *OpenAI[1]*) to generate a MESH search strategy and define the inclusion and exclusion criteria for our review. *We repeated the prompts for MeSH term generation multiple times and refined the MeSH terms by narrowing overly broad terms, incorporating essential terms that were initially omitted by ChatGPT and excluding terms that were not relevant to our review.* Two human researchers, NS and DB, used the MESH terms generated for the PubMed search and retrieved a total of 155 articles. They prepared the articles for presentation to ChatGPT, presenting only the title for title screening, only the abstract for abstract screening, and only the text of the Introduction, Methods, Results, and Discussion for full-text screening, *without any reference to authors or publishing journal, etc*. NS and DB also

created all the prompts for ChatGPT and saved all interactions with the Transformer network. These interactions will be made available as supplementary materials.

Screening Process

Title and abstract screening were conducted independently by ChatGPT 3.5 legacy and the two human researchers, NS and DB. Abstracts were only included for full-text analysis when a consensus was reached between ChatGPT and the human researchers (n=41). NS and DB then generated a table for structured data extraction at the full-text screening level, which will be included in the paper.

Two separate instances of ChatGPT 3.5 legacy were used to independently screen all full texts prepared by NS and DB. NS and DB also evaluated all full-text articles (n=41) for inclusion or exclusion.

Data Extraction and Synthesis

The review of the five included articles was conducted by ChatGPT 4.0 (Version March 15). First, ChatGPT summarized each paper. Next, it was asked to generate a results section and discussion section. All authors extracted data from the included papers and reviewed the text generated by ChatGPT 4.0, making any necessary adjustments and adaptations. *Additionally, tables and charts were generated by human researchers, owing to the constraints of ChatGPT at the time of conducting this study. We extracted the following items in the extraction table (Table S1 in Multimedia Appendix 18): English language (yes/no), targeted disease, study design (randomized controlled trial; cohort study; cross-sectional study; case report or series; meta-analysis, systematic review, or review; opinion; others, experimental, or not applicable), NLP model (yes/no), sample size, performance parameters available (yes/no), clinical task (screening or risk, disease diagnosis, treatment decision, decision support, automation of clinical tasks, data mining or automated document evaluation, others, or not applicable), and clinical implementation (yes/no). The reference directory was compiled by human researchers.*

Statistical Analysis of GPT and Human Performance

In this study, we used several standard performance metrics to evaluate the effectiveness of the search strategy generated by ChatGPT. Below, we describe the calculation of each of these metrics.

Sensitivity (also known as True Positive Rate): Sensitivity is calculated as the number of true positives (TP) divided by the sum of the true positives and the false negatives (FN).

Sensitivity=TP(TP+FN)

Specificity: Specificity is calculated as the number of true negatives (TN) divided by the sum of the true negatives and the false positives (FP).

Specificity=TN(TN+FP)

Precision (also known as Positive Predictive Value): Precision is calculated as the number of true positives (TP) divided by the sum of the true positives and the false positives (FP).

Precision=TP(TP+FP)

Accuracy: Accuracy is calculated as the sum of the true positives (TP) and true negatives (TN) divided by the sum of the true positives, true negatives, false positives, and false negatives.

Accuracy=(TP+TN)(TP+TN+FP+FN)

Chance Hit Rate: The Chance Hit Rate is calculated as the sum of the product of sensitivity and prevalence, and the product of specificity and (1-prevalence).

Chance Hit Rate=(Sensitivity·Prevalence)+(Specificity·(1−Prevalence))

Statistical analysis was carried out by the human researchers due to limitations of ChatGPT at that time. The inter- and intrarater reliability, sensitivity, specificity, and other statistics were calculated using the inclusion/exclusion table created by all authors. The extracted data were compared using Fleiss and Cohen κ, correlating results from both human researchers and 2 iterations of ChatGPT 3.5 [35,36]. Consensus among human researchers was considered the gold standard to compare the performance of the human researchers with that of ChatGPT 3.5 iterations. When a specific measure was not considered as an item, we used a binary categorization of "correct" or "incorrect" as the items (ie, sample size) for assessing inter- and intrarater reliability, decided by consensus with consultation of a third human researcher.

Results

MESH Search and Screening Process

Our MESH search on PubMed yielded 155 papers. Upon screening the titles, the two human researchers included 75 papers, while ChatGPT 3.5 included 115, achieving a sensitivity of [original: 97.33%] *100%* and specificity of [original: 37.5%] *50%(precision=65.2%, accuracy=74.2%, and chance hit rate=49.2%)*. Following the abstract screening of all 155 abstracts, the two human researchers included 41 articles, while ChatGPT 3.5 included 108, resulting in a sensitivity of [original: 95.12] *100%* and specificity of [original: 34.38] 41.2% *(precision=39.6%, accuracy=56.8%, and chance hit rate=40.7%)*. A total of [original: 38] *41* articles were selected for full-text analysis, with 3 articles being excluded due to unavailability. Ultimately, 5 articles were incorporated into our systematic review [37–41].

Natural Language Processing Applications in Various Clinical Settings

Clinical Decision Support System (CDSS) for Concept-Based Searching

Berge et al. developed a machine learning-driven CDSS employing NLP for concept-based searching in a Norwegian hospital [37]. The study introduced an Information System for Clinical Concept-based Search (ICCS) CDSS, devised to detect patient allergies in EHRs using unsupervised machine learning algorithms for clinical narrative analysis. The system combines unsupervised and supervised algorithms with deterministic rules to enhance precision. In a previous study, the ICCS achieved a recall of 92.6%, precision of 88.8%, and F-measure of 90.7%. The ICCS aims to improve allergy detection and classification, thereby enhancing patient safety in anesthesia and ICU settings.

Digital Pathology Applications

Marchesin et al. investigated the use of NLP to strengthen digital pathology applications [38]. The authors introduced explainable knowledge extraction tools capable of extracting pertinent information from pathology reports. They presented the Semantic Knowledge Extractor Tool (SKET), a hybrid knowledge extraction system for digital pathology applications. SKET combines expert knowledge, pre-trained machine learning models, and rule-based techniques such as ScispaCy. The tool exhibits high performance in entity linking and text classification tasks across various cancer use-cases, surpassing unsupervised approaches. The web-based system, SKET X, enables domain experts to understand SKET's outcomes, rules, and parameters for explainable AI. Applications include automatic report annotation, pathological knowledge visualization, and Whole Slide Image classification.

Identification of Nonvalvular Atrial Fibrillation (NVAF)

Elkin et al. employed artificial intelligence with NLP to integrate electronic health record (EHR) structured and free-text data to identify NVAF, aiming to reduce strokes and death [39]. The study utilized high-definition NLP (HD-NLP) to process free text in EHRs, identifying patients with Nonvalvular Atrial Fibrillation (NVAF) and estimating their stroke and bleeding risks. NLP-assisted analysis of structured and unstructured EHR data improved detection rates and accuracy compared to structured data alone. This approach could potentially prevent 176,537 strokes, 10,575 deaths, and save over $18 billion in the first year if implemented nationally, with a net financial benefit of approximately $14.4 billion.

Cardiovascular Disease Comorbidity Assessment

Berman et al. applied NLP to assess cardiovascular disease comorbidities in the Cardio-Canary Comorbidity Project [40]. The authors demonstrated the potential of NLP in facilitating the identification of comorbidities, leading to improved patient care and outcomes in cardiovascular disease management. The modules exhibited robust performance, particularly for hypertension, dyslipidemia, and stroke, with over 95% positive predictive value (PPV) for note-level performance. The NLP modules provide an accurate, open-source system for various applications, such as population management, clinical research, and clinical trial recruitment.

Post-traumatic stress disorder (PTSD) Quality Metrics Improvement

Shiner et al. explored the use of NLP to enhance PTSD quality metrics in psychotherapy treatments for veterans [41]. The study combined structured EMR data with NLP-derived data to evaluate PTSD care quality in the Veteran Affairs system. The validated NLP algorithm displayed a high degree of agreement with template data (weighted kappa: 0.81), capturing nearly 90% of evidence based psychotherapy for PTSD visit days. The study revealed that 20% of PTSD checklist values were documented exclusively in free-text clinical notes. The findings suggest that NLP can bridge documentation gaps, provide a more comprehensive view of care quality, and improve measurement practices for PTSD patients within the Veterans Affairs healthcare system.

Comparison Between ChatGPT and Human Researchers

Except for clinical tasks (κ=0.56), both human researchers showed very good agreement (κ>0.90) for the parameters extracted from the included articles (Table 1). ChatGPT and the human researchers showed very good agreement for the article's language (κ=1), targeted disease (κ=1), NLP model (κ=0.95), sample size (κ=0.83), and performance parameters (κ=0.85); good agreement for study design (κ=0.79); moderate agreement for clinical task (κ=0.58); and only fair agreement for clinical implementation (κ=0.34). All numbers were extracted correctly from the articles by ChatGPT.

In the process of composition, ChatGPT was prompted to provide source citations (refer to Table S2 in Multimedia Appendix 19). Among the 28 references supplied, 3 were found to be fictitious: Smith, Brown & Lee (2022), Demner-Fushman & Chapman (2017), and McCoy, Hughes, Jao & Perlis (2019); this rendered the attribution of Smith, Lee and Joa uncertain. Although the other authors have multiple publications within the NLP domain, a reliable attribution remains elusive. Five of the references were thematically pertinent, yet they did not accurately substantiate the statements made. Additionally, 2 sources required corrections to their publication years. Consequently, a total of 15 references were amended, appended, or substituted.

Table 1. Inter- and intrarater reliability for extraction items using Fleiss and Cohen κ.

	All (Fleiss κ)	GPT[a] 3.5 1 vs 2 (Cohen κ)	Human researcher 1 vs 2 (Cohen κ)
Language	1	1	1
Targeted disease	1	1	1
Study design	0.7333676	0.78939034	0.94736842
NLP[b] model	0.1441441	0.947331947	1
Sample size	–0.041096	0.829723674	0.91486184
Performance	0.6847407	0.853733641	0.9425548
Clinical Task	0.5615047	0.58372457	0.56422018
Implementation	0.3531915	0.34127844	0.92132505

[a]GPT: Generative Pre-Trained Transformer.

[b]NLP: natural language processing.

Discussion

This systematic review aimed to investigate the current natural language processing (NLP) models being used in daily clinical practice. We identified five studies that showcased various applications of NLP in clinical settings, including clinical decision support systems, digital pathology applications, identification of nonvalvular atrial fibrillation, cardiovascular disease comorbidity assessment, and PTSD quality metrics improvement. These studies highlight the potential of NLP to revolutionize healthcare by improving efficiency, accuracy, and patient care.

Berge et al. [37] presented a clinical decision support system (CDSS) that uses NLP for concept-based searching in a Norwegian hospital. Their study demonstrated the potential of machine learning-driven CDSS to improve allergy detection and classification, leading to enhanced patient safety in anesthesia and ICU settings. Marchesin et al. [38] focused on the application of NLP in digital pathology applications, showcasing how NLP can support pathologists and improve the overall quality of pathology diagnosis and patient care. Elkin et al. [39] showed the effectiveness of NLP in identifying NVAF patients, which has the potential to lead to better management of NVAF and prevent strokes and death. Berman et al. [40] utilized NLP for cardiovascular disease comorbidity assessment, illustrating the potential of NLP to facilitate the identification of comorbidities, leading to improved patient care and outcomes. Lastly, Shiner et al. [41] examined the use of NLP to improve PTSD quality metrics in psychotherapy treatments for veterans, demonstrating NLP's value in capturing important data in large healthcare systems and improving measurement practices.

The studies included in this review showcased various NLP techniques, such as machine learning algorithms, rule-based techniques, and the use of pre-trained models like ScispaCy. These approaches demonstrate the versatility of NLP in handling different clinical tasks and highlight the potential for continued development in this field. Moreover, the use of transformer-based models like GPT-3 in conducting this systematic review serves as an example of how NLP can improve the efficiency and accuracy of literature synthesis in a streamlined manner [32].

Despite the promising results, the studies included in this review also have some limitations. First, the studies are limited in terms of the variety of clinical applications and settings, as only five studies were included in the review. This could potentially limit the generalizability of the findings. Furthermore, the studies may have inherent biases and limitations that could impact the interpretation of the results. It is essential to be cautious when extrapolating these findings to other contexts and clinical settings.

Future research should focus on expanding the range of clinical applications and settings where NLP can be utilized, as well as investigating the scalability and generalizability of the identified approaches. Additionally, more studies should be conducted to explore the potential of transformer-based models like GPT-3 and BERT in clinical practice. These models have shown great promise in various NLP tasks and may offer further advancements in the field of healthcare.

In conclusion, our systematic review highlights the potential of NLP in revolutionizing clinical practice by improving efficiency, accuracy, and patient care. The studies included in this review showcase various NLP applications in clinical settings, demonstrating the versatility and potential for growth in this field. Further research is needed to expand the range of clinical applications and settings, as well as to explore the potential of transformer-based models in healthcare. As NLP continues to advance, it is expected that its impact on clinical practice will only increase, leading to improved patient outcomes and more efficient healthcare systems.

Concluding Remarks by the Human Authors

Concerning the systematic review, we only searched PubMed and no other database or registry. Furthermore, the MeSH search generated only 155 hits, and we must admit that this study does not allow us to determine whether NLPs are of practical use in clinical practice today. Since the MeSH term itself was produced by ChatGPT and the main goal of this study was to explore the usefulness of ChatGPT in performing or assisting in systematic reviews, we adhered to the generated methods; however, this compromises the quality of the

systematic review. Therefore, we do not believe that an adequate commentary on the state and usefulness of NLP in clinical practice is within the scope of this study. During the research, ChatGPT underwent several updates. We attempted to split workflows for both GPT 3.5 and GPT 4.0. Since developments on LLMs change at a fast pace, their applicability might change fast as well, which means that results from interactions and the idea of augmented or automated systematic reviews can change drastically, for example, with developments in LLMs' ability to access of databases such as PubMed or Cochrane.

We hypothesize that automated systematic reviews could become a reality in the near future. However, the current state of ChatGPT versions 3.5 and 4.0, with their multiple limitations, renders augmented systematic reviews inefficient for experienced researchers. Yet, for language correction, particularly for nonnative English speakers, and rectification of grammatical errors, or for text condensation and modification in form and wording, it proves to be of significant value. As we confined our study to ChatGPT, without the use of any plugins or application programming interface implementations, we anticipate that the forthcoming months or years will witness an increased application of LLMs in scientific research, showcasing intriguing architectures such as "Lang Chain" and "Agent GPT" as pioneering examples of more complex programs powered by LLMs.

Ethical and legal concerns about the implementation of LLMs in a scientific field as sensible as medicine have led to an ongoing discussion and should be considered before broadening the spectrum of clinical applications for NLP-driven automations. The black box issue associated with LLMs such as ChatGPT, even when using the most deterministic options, is an undeniable fact. Automatic analyses of all available literature within minutes or seconds, however, would change the way we conduct research or are able to access information in clinical practice. Further research is imperative, accompanied by a debate on the ethical implications of such potent tools and strategies to oversee and regulate the use of these models in scientific writing.

Acknowledgments

We would like to thank the researchers and authors who contributed to the studies included in this systematic review. Several parts of the text, including the original draft of the abstract, were generated using ChatGPT (versions 3.5 and 4.0, OpenAI [1]) and have been indicated as such within the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Article sumarization.
[TXT File , 279 KB - medinform_v11i1e48933_app1.txt]

Multimedia Appendix 2
Article sumarization.
[TXT File , 1099 KB - medinform_v11i1e48933_app2.txt]

Multimedia Appendix 3
Data extraction.
[TXT File , 93 KB - medinform_v11i1e48933_app3.txt]

Multimedia Appendix 4
Extraction fulltext.
[TXT File , 1063 KB - medinform_v11i1e48933_app4.txt]

Multimedia Appendix 5
Full text analysis GPT1.
[TXT File , 1098 KB - medinform_v11i1e48933_app5.txt]

Multimedia Appendix 6
Full text analysis GPT2.
[TXT File , 459 KB - medinform_v11i1e48933_app6.txt]

Multimedia Appendix 7
MESH-Term inclusion/exclusion criteria.

[TXT File , 60 KB - medinform_v11i1e48933_app7.txt]

Multimedia Appendix 8
Title screening GPT1.
[TXT File , 65 KB - medinform_v11i1e48933_app8.txt]

Multimedia Appendix 9
Title screening GPT2.
[TXT File , 497 KB - medinform_v11i1e48933_app9.txt]

Multimedia Appendix 10
Abstract screening GPT1.
[TXT File , 43 KB - medinform_v11i1e48933_app10.txt]

Multimedia Appendix 11
Writing abstract and title.
[TXT File , 46 KB - medinform_v11i1e48933_app11.txt]

Multimedia Appendix 12
Writing methods.
[TXT File , 18 KB - medinform_v11i1e48933_app12.txt]

Multimedia Appendix 13
Writing results and conclusion.
[TXT File , 11 KB - medinform_v11i1e48933_app13.txt]

Multimedia Appendix 14
Rewrite results.
[TXT File , 26 KB - medinform_v11i1e48933_app14.txt]

Multimedia Appendix 15
Supplementary Materials, Task and Prompts correlated.
[XLSX File (Microsoft Excel File), 12 KB - medinform_v11i1e48933_app15.xlsx]

Multimedia Appendix 16
Supplementary Materials, Task and Prompts correlated.
[XLSX File (Microsoft Excel File), 12 KB - medinform_v11i1e48933_app16.xlsx]

Multimedia Appendix 17
PRISMA 2020 Checklist.
[DOCX File , 32 KB - medinform_v11i1e48933_app17.docx]

Multimedia Appendix 18
Two human researchers (NS, DB) and two instances of ChatGPT (GPT1, GPT2) were tasked with individually checking the full text of the articles for inclusion and exclusion criteria, as well as extracting data endpoints.
[XLSX File (Microsoft Excel File), 15 KB - medinform_v11i1e48933_app18.xlsx]

Multimedia Appendix 19
Control of the references.
[XLSX File (Microsoft Excel File), 14 KB - medinform_v11i1e48933_app19.xlsx]

References

1. OpenAI. URL: https://openai.com/ [accessed 2023-10-18]
2. Jiang M, Chen Y, Liu M, Rosenbloom ST, Mani S, Denny JC, et al. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. J Am Med Inform Assoc 2011 Sep 01;18(5):601-606 [FREE Full text] [doi: 10.1136/amiajnl-2011-000163] [Medline: 21508414]

3. Haug C, Drazen J. Artificial intelligence and machine learning in clinical medicine, 2023. N Engl J Med 2023 Mar 30;388(13):1201-1208 [FREE Full text] [doi: 10.1056/nejmra2302038]
4. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform 2018 Mar 07;17(01):128-144. [doi: 10.1055/s-0038-1638592]
5. Rajpurkar P, Chen E, Banerjee O, Topol E. AI in health and medicine. Nat Med 2022 Jan;28(1):31-38 [FREE Full text] [doi: 10.1038/s41591-021-01614-0] [Medline: 35058619]
6. Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, et al. Clinical information extraction applications: a literature review. J Biomed Inform 2018 Jan;77:34-49. [doi: 10.1016/j.jbi.2017.11.011] [Medline: 29162496]
7. Shivade C, Raghavan P, Fosler-Lussier E, Embi PJ, Elhadad N, Johnson SB, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. J Am Med Inform Assoc 2014 Mar 01;21(2):221-230 [FREE Full text] [doi: 10.1136/amiajnl-2013-001935] [Medline: 24201027]
8. Lederman A, Lederman R, Verspoor K. Tasks as needs: reframing the paradigm of clinical natural language processing research for real-world decision support. J Am Med Inform Assoc 2022 Sep 12;29(10):1810-1817 [FREE Full text] [doi: 10.1093/jamia/ocac121] [Medline: 35848784]
9. Reddy V, Nafees A, Raman S. Recent advances in artificial intelligence applications for supportive and palliative care in cancer patients. Curr Opin Support Palliat Care 2023 Jun 01;17(2):125-134. [doi: 10.1097/SPC.0000000000000645] [Medline: 37039590]
10. Odisho AY, Bridge M, Webb M, Ameli N, Eapen RS, Stauf F, et al. Automating the capture of structured pathology data for prostate cancer clinical care and research. JCO Clin Cancer Inform 2019 Dec(3):1-8. [doi: 10.1200/cci.18.00084]
11. Feng Y, Liang S, Zhang Y, Chen S, Wang Q, Huang T, et al. Automated medical literature screening using artificial intelligence: a systematic review and meta-analysis. J Am Med Inform Assoc 2022 Jul 12;29(8):1425-1432 [FREE Full text] [doi: 10.1093/jamia/ocac066] [Medline: 35641139]
12. Devlin J, Chang M, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv. Preprint posted online October 11, 2018 [FREE Full text]
13. Brown T, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. arXiv. Preprint posted online May 28, 2020 [FREE Full text]
14. Lee J, Yoon W, Kim S, Kim D, Kim S, So C, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020 Feb 15;36(4):1234-1240 [FREE Full text] [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
15. Alsentzer E, Murphy J, Boag W, Weng W, Jindi D, Naumann T, et al. Publicly Available Clinical BERT Embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop. 2019 Presented at: 2nd Clinical Natural Language Processing Workshop; June 2019; Minneapolis, MN p. 72-78. [doi: 10.18653/v1/w19-1909]
16. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A, et al. Attention is All you Need. 2017 Presented at: 31st Conference on Neural Information Processing Systems; 2017; Long Beach, CA.
17. Strubell E, Ganesh A, McCallum A. Energy and policy considerations for deep learning in NLP. arXiv. Preprint posted online June 5, 2019 [FREE Full text] [doi: 10.18653/v1/p19-1355]
18. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of hallucination in natural language generation. ACM Comput Surv 2023 Mar 03;55(12):1-38 [FREE Full text] [doi: 10.1145/3571730]
19. Weng W, Wagholikar K, McCray A, Szolovits P, Chueh H. Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach. BMC Med Inform Decis Mak 2017 Dec 01;17(1):155 [FREE Full text] [doi: 10.1186/s12911-017-0556-8] [Medline: 29191207]
20. Beam AL, Kohane IS. Big data and machine learning in health care. JAMA 2018 Apr 03;319(13):1317-1318. [doi: 10.1001/jama.2017.18391] [Medline: 29532063]
21. Fu S, Chen D, He H, Liu S, Moon S, Peterson K, et al. Clinical concept extraction: a methodology review. J Biomed Inform 2020 Sep;109:103526 [FREE Full text] [doi: 10.1016/j.jbi.2020.103526] [Medline: 32768446]
22. Yan M, Gustad L, Nytrø Ø. Sepsis prediction, early detection, and identification using clinical text for machine learning: a systematic review. J Am Med Inform Assoc 2022 Jan 29;29(3):559-575 [FREE Full text] [doi: 10.1093/jamia/ocab236] [Medline: 34897469]
23. Manning CD. Daedalus 2022;151(2):127-138 [FREE Full text] [doi: 10.1162/daed_a_01905]
24. Moher D, Liberati A, Tetzlaff J, Altman D, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 2009 Jul 21;6(7):e1000097 [FREE Full text] [doi: 10.1371/journal.pmed.1000097] [Medline: 19621072]
25. Page M, McKenzie J, Bossuyt P, Boutron I, Hoffmann T, Mulrow C, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
26. Bramer W, Giustini D, de Jonge GB, Holland L, Bekhuis T. De-duplication of database search results for systematic reviews in EndNote. J Med Libr Assoc 2016 Jul;104(3):240-243 [FREE Full text] [doi: 10.3163/1536-5050.104.3.014] [Medline: 27366130]

27. Gøtzsche PC, Ioannidis J. Content area experts as authors: helpful or harmful for systematic reviews and meta-analyses? BMJ 2012 Nov 01;345:e7031 [FREE Full text] [doi: 10.1136/bmj.e7031] [Medline: 23118303]
28. Larsen P, von Ins M. The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. Scientometrics 2010 Sep;84(3):575-603 [FREE Full text] [doi: 10.1007/s11192-010-0202-z] [Medline: 20700371]
29. Waffenschmidt S, Janzen T, Hausner E, Kaiser T. Simple search techniques in PubMed are potentially suitable for evaluating the completeness of systematic reviews. J Clin Epidemiol 2013 Jun;66(6):660-665 [FREE Full text] [doi: 10.1016/j.jclinepi.2012.11.011] [Medline: 23419611]
30. Snyder H. Literature review as a research methodology: an overview and guidelines. J Bus Res 2019 Nov;104:333-339 [FREE Full text] [doi: 10.1016/j.jbusres.2019.07.039]
31. Tsafnat G, Glasziou P, Choong M, Dunn A, Galgani F, Coiera E. Systematic review automation technologies. Syst Rev 2014 Jul 09;3:74 [FREE Full text] [doi: 10.1186/2046-4053-3-74] [Medline: 25005128]
32. Marshall I, Wallace B. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. Syst Rev 2019 Jul 11;8(1):163 [FREE Full text] [doi: 10.1186/s13643-019-1074-9] [Medline: 31296265]
33. Higgins J, Green S. Cochrane Handbook for Systematic Reviews of Interventions: Cochrane Book Series. London: The Cochrane Collaboration; 2008.
34. Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al. Cochrane Handbook for Systematic Reviews of Interventions version 6.3. London: The Cochrane Collaboration; 2022.
35. Fleiss J. Measuring nominal scale agreement among many raters. Psychological Bulletin 1971 Nov;76(5):378-382 [FREE Full text] [doi: 10.1037/h0031619]
36. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 2016 Jul 02;20(1):37-46 [FREE Full text] [doi: 10.1177/001316446002000104]
37. Berge G, Granmo O, Tveit T, Munkvold B, Ruthjersen A, Sharma J. Machine learning-driven clinical decision support system for concept-based searching: a field trial in a Norwegian hospital. BMC Med Inform Decis Mak 2023 Jan 10;23(1):5 [FREE Full text] [doi: 10.1186/s12911-023-02101-x] [Medline: 36627624]
38. Marchesin S, Giachelle F, Marini N, Atzori M, Boytcheva S, Buttafuoco G, et al. Empowering digital pathology applications through explainable knowledge extraction tools. J Pathol Inform 2022;13:100139 [FREE Full text] [doi: 10.1016/j.jpi.2022.100139] [Medline: 36268087]
39. Elkin P, Mullin S, Mardekian J, Crowner C, Sakilay S, Sinha S, et al. Using artificial intelligence with natural language processing to combine electronic health record's structured and free text data to identify nonvalvular atrial fibrillation to decrease strokes and death: evaluation and case-control study. J Med Internet Res 2021 Nov 09;23(11):e28946 [FREE Full text] [doi: 10.2196/28946] [Medline: 34751659]
40. Berman A, Biery D, Ginder C, Hulme O, Marcusa D, Leiva O, et al. Natural language processing for the assessment of cardiovascular disease comorbidities: The cardio-Canary comorbidity project. Clin Cardiol 2021 Sep;44(9):1296-1304 [FREE Full text] [doi: 10.1002/clc.23687] [Medline: 34347314]
41. Shiner B, Levis M, Dufort V, Patterson OV, Watts BV, DuVall SL, et al. Improvements to PTSD quality metrics with natural language processing. J Eval Clin Pract 2022 Aug;28(4):520-530 [FREE Full text] [doi: 10.1111/jep.13587] [Medline: 34028937]

Abbreviations

BERT: Bidirectional Encoder Representations from Transformers
CDSS: clinical decision support system
EHR: electronic health record
FN: false negative
FP: false positive
GPT: Generative Pre-trained Transformer
ICCS: Information System for Clinical Concept-based Search
LLM: large language model
MeSH: Medial Subject Headings
NLP: natural language processing
NVAF: nonvalvular atrial fibrillation
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PTSD: posttraumatic stress disorder
SKET: Semantic Knowledge Extractor Tool
TN: true negative
TP: true positive

Edited by C Lovis; submitted 11.05.23; peer-reviewed by V Ochs, T Hou; comments to author 31.05.23; revised version received 20.06.23; accepted 25.08.23; published 28.11.23.

Please cite as:
Schopow N, Osterhoff G, Baur D
Applications of the Natural Language Processing Tool ChatGPT in Clinical Practice: Comparative Study and Augmented Systematic Review
JMIR Med Inform 2023;11:e48933
URL: https://medinform.jmir.org/2023/1/e48933
doi:10.2196/48933
PMID:38015610

Original Paper

Risk Prediction of Emergency Department Visits in Patients With Lung Cancer Using Machine Learning: Retrospective Observational Study

Ah Ra Lee[1], PhD; Hojoon Park[1], BSB; Aram Yoo[1], BSN, RN; Seok Kim[1], MPH; Leonard Sunwoo[2], MD, PhD; Sooyoung Yoo[1], PhD

[1]Office of eHealth Research and Business, Seoul National University Bundang Hospital, Seongnam-si, Republic of Korea

[2]Department of Radiology, Seoul National University Bundang Hospital, Seongnam-si, Republic of Korea

Corresponding Author:
Sooyoung Yoo, PhD
Office of eHealth Research and Business
Seoul National University Bundang Hospital
172, Dolma-ro, Bundang-gu
Seongnam-si, 13605
Republic of Korea
Phone: 82 31 787 8980
Fax: 82 31 787 4061
Email: yoosoo0@snubh.org

Abstract

Background: Patients with lung cancer are among the most frequent visitors to emergency departments due to cancer-related problems, and the prognosis for those who seek emergency care is dismal. Given that patients with lung cancer frequently visit health care facilities for treatment or follow-up, the ability to predict emergency department visits based on clinical information gleaned from their routine visits would enhance hospital resource utilization and patient outcomes.

Objective: This study proposed a machine learning–based prediction model to identify risk factors for emergency department visits by patients with lung cancer.

Methods: This was a retrospective observational study of patients with lung cancer diagnosed at Seoul National University Bundang Hospital, a tertiary general hospital in South Korea, between January 2010 and December 2017. The primary outcome was an emergency department visit within 30 days of an outpatient visit. This study developed a machine learning–based prediction model using a common data model. In addition, the importance of features that influenced the decision-making of the model output was analyzed to identify significant clinical factors.

Results: The model with the best performance demonstrated an area under the receiver operating characteristic curve of 0.73 in its ability to predict the attendance of patients with lung cancer in emergency departments. The frequency of recent visits to the emergency department and several laboratory test results that are typically collected during cancer treatment follow-up visits were revealed as influencing factors for the model output.

Conclusions: This study developed a machine learning–based risk prediction model using a common data model and identified influencing factors for emergency department visits by patients with lung cancer. The predictive model contributes to the efficiency of resource utilization and health care service quality by facilitating the identification and early intervention of high-risk patients. This study demonstrated the possibility of collaborative research among different institutions using the common data model for precision medicine in lung cancer.

(JMIR Med Inform 2023;11:e53058) doi:10.2196/53058

KEYWORDS

emergency department; lung cancer; risk prediction; machine learning; common data model; emergency; hospitalization; hospitalizations; lung; cancer; oncology; lungs; pulmonary; respiratory; predict; prediction; predictions; predictive; algorithm; algorithms; risk; risks; model; models

Introduction

Lung cancer is a well-known malignancy that causes severe respiratory symptoms. There were 1.8 million lung cancer–related fatalities and 2.2 million newly diagnosed patients worldwide in 2020 [1]. Patients with lung cancer frequently encounter complex health care challenges, such as unanticipated visits to the emergency department (ED) due to disease progression, treatment-related problems, and comorbidities [2]. Despite advances in medical technology, which have increased the survival rates of patients with lung cancer, ongoing management is still needed after initial oncology treatment due to the diverse characteristics and causes of the disease and the fact that each patient's disease stage and conditions vary [3].

Patients with lung cancer often experience acute complications or disease progression that may require urgent medical attention [4]. The frequency of ED visits among patients with lung cancer increases with the length of their survival [5]. Prior research has shown that approximately 10% of all cancer-related ED visits are attributable to lung cancer [6]. In addition, compared to patients with other types of cancer, those with lung cancer who visit the ED tend to have a worse prognosis [7]. Shin et al [8] examined the 28-day mortality rate among intubated patients with cancer in the ED. Their findings revealed that patients with lung cancer faced a higher mortality risk compared to those with other types of cancer. Another previous study found that patients with lung cancer had the highest mortality rate (48.1%) within 28 days in contrast to those with other cancers who presented to the ED with septic shock [9]. Further, a comprehensive study conducted on hospitalizations related to sepsis incidence and mortality rates among patients with cancer in the United States revealed that patients with lung cancer had the highest mortality rate [10]. These findings from previous studies demonstrate the poor prognosis for patients with severe lung cancer–related conditions in the ED.

To lower the risk of a poor prognosis, it is of the utmost importance to predict visits to the ED among patients with lung cancer in advance [11]. Nevertheless, the existing literature on this subject is limited, with only a handful of studies identifying the risk factors associated with these visits. Consequently, the lack of comprehensive research in this field hinders clinicians' ability to intervene in a timely manner. Hong et al [12] developed a machine learning–based model for predicting ED visits in patients who were undergoing radiotherapy or chemoradiotherapy. Sutradhar et al [13] proposed a risk prediction approach for ED visits among patients with cancer using the Edmonton symptom assessment system and conventional statistical analysis and logistic regression methods [13]. Sutradhar and Barbera [14] also developed a machine learning model to predict 7-day ED visits in patients with cancer using the Edmonton symptom assessment system and other clinical information. Previous studies were predominantly conducted among patients with all types of cancer, so they did not identify risk factors specific to those with lung cancer for ED visits.

Therefore, the primary objective of this study was to identify the influential factors for patients with lung cancer that may impact ED visits. This study developed a machine learning–based prediction model to forecast ED visits among patients with lung cancer using data from the Observational Medical Outcomes Partnership (OMOP) common data model (CDM) [15]. Additionally, the significance of various types of clinical information in influencing the decision-making process of the machine learning model was evaluated. The ability to predict ED visits enables health care service providers to identify high-risk patients in advance, allowing for prompt intervention and appropriate management. By intervening expeditiously, clinicians may be able to prevent or mitigate emergency situations, leading to improved patient outcomes. This study contributes to minimizing preventable health deterioration by facilitating early intervention during lung cancer treatment by clinicians.

Methods

Study Design and Source of Data

This was a retrospective observational study using electronic health records at Seoul National University Bundang Hospital (SNUBH), a tertiary general hospital in South Korea. The electronic health record data were converted to the OMOP CDM, which is a standardized data format in the observational health data sciences and informatics community. The data set comprised information from visit histories encompassing a broad range of categories, such as diagnoses, medication prescriptions, laboratory test results, performed procedures, and clinical observations.

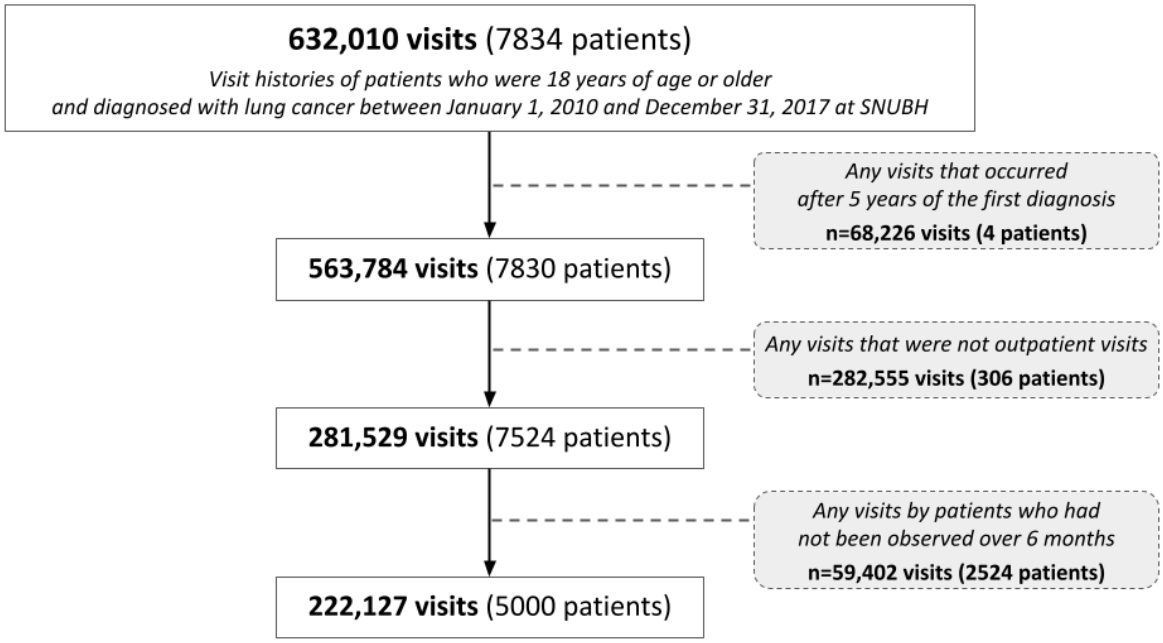
Ethical Considerations

This study was conducted with approval and waivers of informed consent or exemptions from the SNUBH institutional review board (no. X-2308-844-903). The CDM data used in this study were deidentified and are securely maintained within an internal network.

Target Population

Patients who were 18 years of age or older and diagnosed with lung cancer (International Classification of Diseases, Tenth Revision code C34: malignant neoplasm of bronchus and lung [16]) at least once between 2010 and 2017 were eligible. Figure 1 presents the inclusion and exclusion criteria for the study participants. This study included all outpatient visits within 5 years of the patient's initial diagnosis of lung cancer. Moreover, since the focus of this study was on health information routinely collected in hospitals based on the CDM, target populations should visit the hospital regularly for disease management. Considering the follow-up period in the treatment guidelines for patients with lung cancer, patients who were unable to follow up for more than 6 months during the study period were excluded.

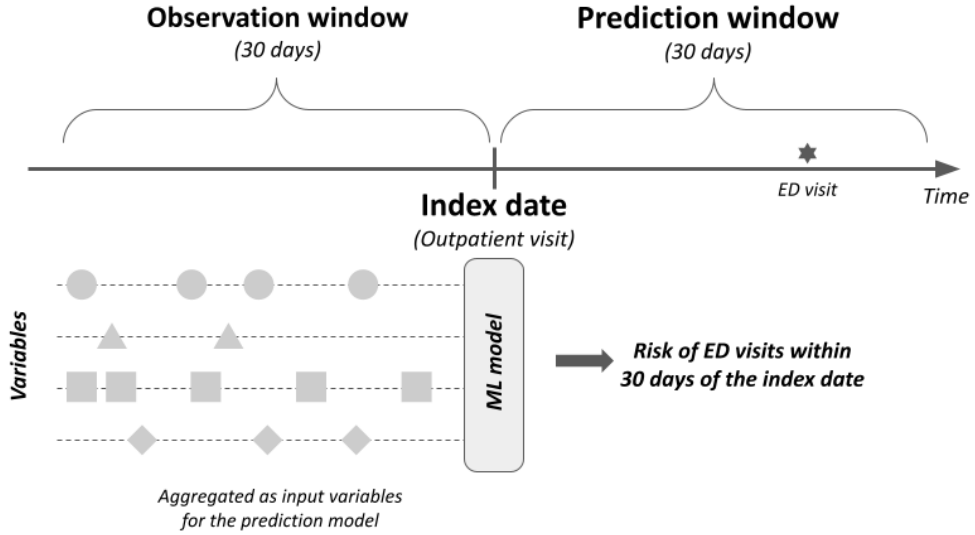
Figure 1. Flowchart of the study participant selection process. SNUBH: Seoul National University Bundang Hospital.



Primary Outcome

Figure 2 presents an overview of the observation and prediction windows for the risk prediction task. The outpatient visits were defined as the index date. Multiple outpatient visits for the same patient were considered independent index dates. During the 30 days prior to the index date, predictors were extracted and aggregated to be used as input variables for the machine learning models. The primary outcome was the occurrence of ED visits within the 30-day period following the index dates. Since the objective of this study was to predict preventable ED visits that occurred during disease progression or treatment among patients with lung cancer, the primary outcome was defined as ED visits that satisfied specific criteria. Valid outcomes were restricted to ED visits that did not result in a transfer to another hospital.

Figure 2. An overview of the observation and prediction windows for the risk prediction task. ED: emergency department; ML: machine learning.



Predictors

Data extracted from the CDM over successive time windows, referred to as the observation window, for 30 days prior to the index dates were used for candidate predictors. This included patient demographics, visit histories, clinical information, laboratory results, and vital signs. If patients had multiple records for the same data item during the observation window, the median value was calculated. Detailed information on the selected features and their data sources is presented in Multimedia Appendix 1.

The selected features used to predict the primary outcome were analyzed using statistical methods. This study provided descriptive statistics for continuous variables in the form of median and IQR values, whereas categorical variables were presented as frequencies and respective percentages. The *P* value of each variable was also calculated to explore the probability of a relationship between the selected features for

prediction tasks. In the significance test, the Mann-Whitney U test was used when analyzing continuous variables, while the chi-square test was used for categorical data.

Model Training and Evaluation

This study defined the prediction task for each event date as a binary classification problem. We used 4 different machine learning models—logistic regression, random forest, extreme gradient boosting, and light gradient boosting machine (LGBM)—to discover the model with the highest performance. These models were selected to provide a comparison of models by evaluating the performances of various alternatives, ranging from conventional linear-based approaches to more complex methods, such as ensemble-based models.

The selected features were preprocessed for use as input variables in the machine learning models. Missing values were replaced with their median value, and aberrant observations that were not acceptable from a theoretical perspective were removed based on prior research and the expertise of clinical domain experts to prevent extreme outliers from leading to failure in the prediction task. The entire data set was split into training and testing sets using repeated 7-fold cross-validation, with stratified sampling accounting for the incidence of the primary outcome. Demographic information of the patients, including age, gender, and comorbidities, was also taken into account given that some patients had made multiple outpatient visits, which corresponded to the index dates in this study. By using repeated k-fold cross-validation, the estimated performance of a machine learning model can be enhanced. The cross-validation procedure was iterated 1000 times. Finally, categorical variables were encoded, and continuous variables were normalized.

The following 4 performance metrics were used for model evaluation: area under the receiver operating characteristic curve (AUROC), area under the precision-recall curve (PRAUC), sensitivity, and specificity. The evaluation results were reported as the average value encompassing all folds from all iterations, and the 95% CI was estimated. This approach ensures a more accurate estimate of the true unknown underlying mean performance of the model on the data set than using the standard error. Using the Shapley additive explanations (SHAP) value from the best performing model, the significance of the features was also analyzed to identify risk factors for ED visits by patients with lung cancer [17]. All the experiments were performed using the Python 3.7.6 environment (Python Software Foundation).

Results

Baseline Characteristics

A total of 222,127 outpatient visits occurred for 5,000 patients. Table 1 presents the baseline characteristics of the target population. The median age was 67.23 (IQR 60.54-75.57), and 3812 (76.24%) patients were older than 60 years. Men accounted for 64.14% (n=3322) of the participants, which was slightly greater than the number of women (n=1678). According to the Charlson comorbidity index (CCI) [18], more than 66.4% (n=3320) of participants did not have any comorbidities before being diagnosed with lung cancer. There were 1212 (24.24%) patients who had a history of smoking, while the others had no smoking history or their history was unknown. About half (n=2827, 56.54%) of the patients lived in the metropolitan area near the hospital.

Table 1. Baseline characteristics of the selected participants (n=5000).

Characteristic	Value
Age (years), median (IQR)	67.23 (60.54-75.57)
Age group (years), n (%)	
18-49	376 (7.52)
50-59	812 (16.24)
60-69	1511 (30.22)
≥70	2301 (46.02)
Gender, n (%)	
Men	3322 (64.14)
Women	1678 (35.86)
Body mass index, median (IQR)	22.90 (20.59-24.97)
Charlson comorbidity index, median (IQR)	2.95 (2.00-3.00)
Charlson comorbidity index group, n (%)	
0	3320 (66.4)
1-2	982 (19.64)
3-4	299 (5.98)
≥5	399 (7.98)
Smoking history, n (%)	
Yes	1212 (24.24)
No	957 (19.14)
Unknown	4600 (56.62)
Residence, n (%)	
Greater Seoul	2827 (56.54)
Other[a]	1071 (21.42)
Unknown	1102 (22.04)

[a]Other areas of residence included Gangwon-do, Chungcheongbuk-do, Chungcheongnam-do, Gyeongsangbuk-do, Gyeongsangnam-do, Jeollabuk-do, Jeollanam-do, Jeju-do, Daejeon, Sejong-si, Daegu, Ulsan, Busan, and Gwangju.

There were 8192 visits to the ED, and 2790 patients (55.80% of total patients) had ED visits during the course of their disease. The results of an explanatory data analysis of ED visits by the target population are displayed in Table 2. Of all the visits, 81.33% (n=6663) were from home, while the rest were from outpatient visits or other institutions, including transfers from hospitals, independent clinics, and inpatient care units. Most ED visits occurred between 7 AM and 10 PM. The median length of time spent in the ED was approximately 13.64 (IQR 2.98-15.84) hours. More than half (n=4956) of the patients were discharged home, while the remaining patients were hospitalized or died in hospital. Of all ED visits, 38.32% (n=3139) resulted in hospitalization, and 1.18% (n=97) resulted in death. The most frequent causes of visits to the ED were neoplasms, including malignant neoplasms of the bronchus and lung and secondary malignant neoplasms of other sites. The other primary causes were diseases of the respiratory system. Patients with symptoms, signs, and abnormal clinical and laboratory findings not elsewhere classified primarily visited the ED for the following reasons: fever of other and unknown origin, hemorrhage from respiratory passages, abnormalities of breathing, pain in the throat and chest, and abdominal and pelvic pain.

Table 2. Explanatory data analysis results of emergency department (ED) visits (n=8192) by selected patients.

Feature	Value
Age (years), median (IQR)	67.10 (60.26-75.50)
Gender, n (%)	
Men	5502 (67.16)
Women	2690 (32.84)
Sourced from, n (%)	
Home	6663 (81.33)
Outpatient visits	697 (8.51)
Other institutions	832 (10.16)
Discharged to, n (%)	
Home	4956 (60.50)
Hospitalization	3139 (38.32)
Death	97 (1.18)
Visit time of day, n (%)	
7AM to 3 PM	4842 (59.11)
3 PM to 10 PM	2374 (28.98)
10 PM to 7 AM	976 (11.91)
Time spent in ED, median (IQR)	13.64 (2.98-15.84)
Primary diagnosis (ICD-10[a] codes), n (%)	
Neoplasms (C00-D48)	4242 (51.78)
Diseases of the respiratory system (J00-J99)	967 (11.8)
Symptoms, signs, and abnormal clinical and laboratory findings not elsewhere classified (R00-R99)	886 (10.82)
Diseases of the digestive system (K00-K93)	361 (4.41)
Diseases of the circulatory system (I00-I99)	307 (3.75)
Other	1429 (17.44)

[a]ICD-10: International Classification of Diseases, Tenth Revision.

The complete list of descriptive statistics (N=222,127) of the selected features used as input variables for the predictive model are displayed in Multimedia Appendix 2. The median value of elapsed days since the first diagnosis of lung cancer was about 333 (IQR 103.60-810.42) days. Chemotherapy was administered at 39.27% (n=87,221) of visits, compared to 16.47% (n=36,575) for radiation therapy and 2.32% (n=5159) for lung cancer–related surgery. Analgesics were administered at 27.35% (n=60,744) of visits, and the use of antibacterials for systemic use accounted for 18.83% (n=41,825) of visits. The results of blood tests revealed a median value of 6.27 (IQR 4.87-8.06) for leukocytes, 237.00 (IQR 188.00-296.00) for platelets, 61.90 (IQR 53.50-70.05) for neutrophils, and 12.10 (IQR 10.80-13.30) for hemoglobin. The shock index, which refers to the ratio of the heart rate to systolic blood pressure, was calculated from the collected vital signs and showed a median value of 0.67 (IQR 0.59-0.78).

Performance Evaluation Results of the Machine Learning Models

The performance evaluation results from the machine learning models are presented in Table 3. The overall AUROC score was ≥0.70 in all models, ranging from 0.70 to 0.73. The optimal prediction threshold was defined using the precision-recall curve since the data set had an imbalanced class distribution. The highest PRAUC score was 0.24 in the LGBM model.

Table 3. Performance comparison of the machine learning models.

Model	Sensitivity (95% CI)	Specificity (95% CI)	AUROC[a] (95% CI)	PRAUC[b] (95% CI)
LR[c]	0.76 (0.7554-0.7581)	0.54 (0.5369-0.5398)	0.71 (0.7054-0.7065)	0.22 (0.2191-0.2205)
RF[d]	0.74 (0.7360-0.7395)	0.57 (0.5634-0.5667)	0.71 (0.7086-0.7097)	0.22 (0.2205-0.2221)
XGB[e]	0.77 (0.7729-0.7759)	0.51 (0.5080-0.5112)	0.70 (0.7022-0.7033)	0.21 (0.2134-0.2145)
LGBM[f]	0.76 (0.7575-0.7605)	0.58 (0.5752-0.5780)	0.73 (0.7312-0.7323)	0.24 (0.2360-0.2374)

[a]AUROC: area under the receiver operating characteristic curve.

[b]PRAUC: area under the precision-recall curve.

[c]LR: logistic regression.

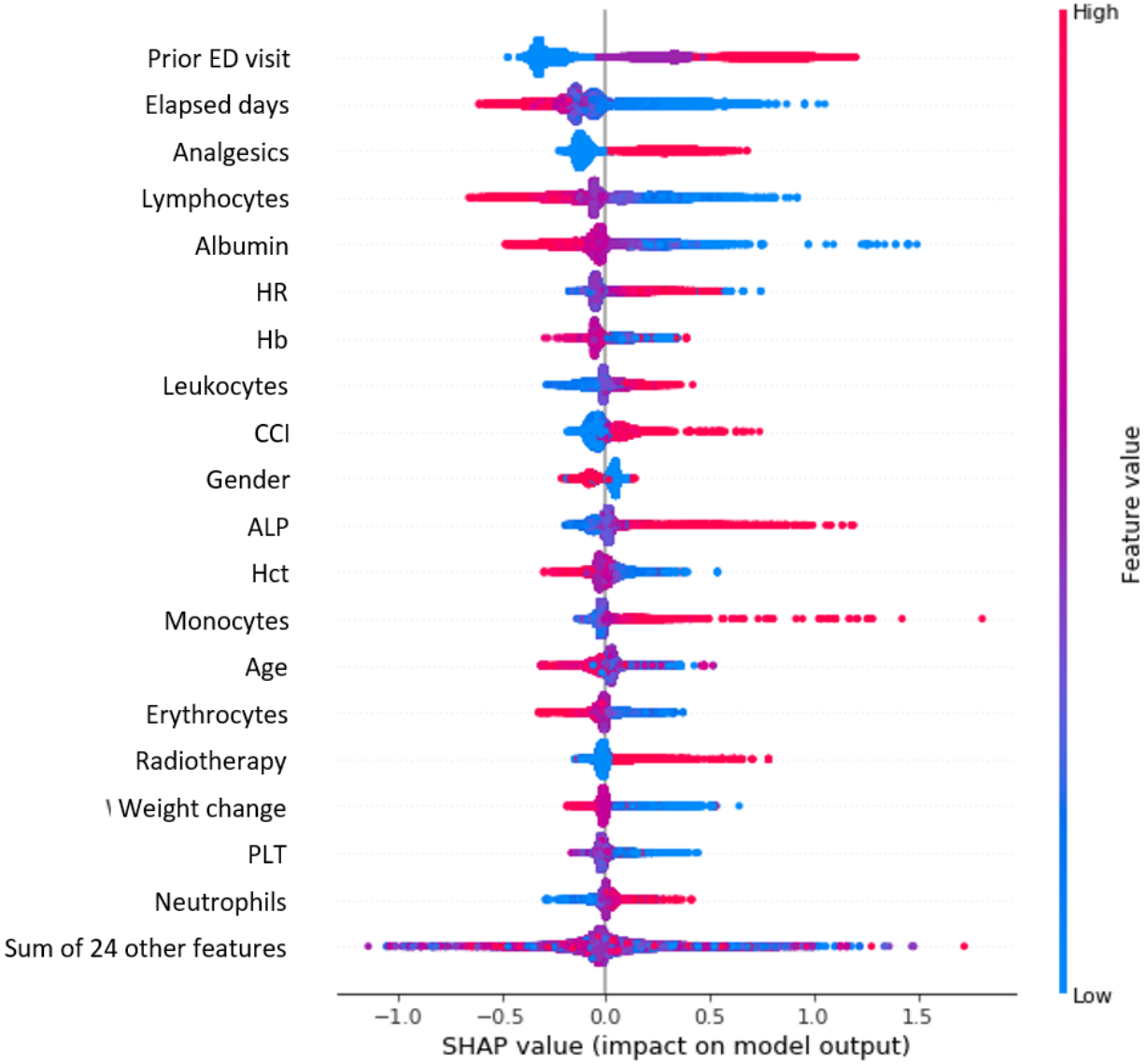
[d]RF: random forest.

[e]XGB: extreme gradient boosting.

[f]LGBM: light gradient boosting machine.

The SHAP values for the highest-performing model, the LGBM, are depicted in Figure 3. The key influencing features for predicting the risk of ED visits in the LGBM model were identified as recent ED visits, elapsed days since the initial diagnosis for lung cancer, the use of analgesics, and lymphocyte and albumin levels. The SHAP value quantifies the influence of a specific feature on the predictions; consequently, a comprehensive interpretation can be obtained by calculating the SHAP values for the model's output. As shown in Figure 3, the number of ED visits, CCI, administration of analgesics, and radiotherapy, as well as high values of leukocytes, alkaline phosphatase, monocytes, and neutrophils, increased the possibility of visits to the ED. In contrast, the combined presence of low values of lymphocytes, albumin, hemoglobin, and hematocrit indicated an increased likelihood of visiting the ED.

Figure 3. Summary plots for SHAP values in the light gradient boosting machine model. ALP: alkaline phosphatase; CCI: Charlson comorbidity index; ED: emergency department; Hb: hemoglobin; Hct: hematocrit; HR: heart rate; PLT: platelets; SHAP: Shapley additive explanations.



Discussion

This study developed a predictive model for ED visits among patients with lung cancer using machine learning and CDM data. Predicting ED visits among patients with lung cancer has numerous important implications. First, the ED visit prediction model is able to facilitate care coordination and patient management. As shown in Table 2, over 38% of all visits resulted in hospitalization. These results were marginally lower than the findings in prior literature, which indicated that 54.5% of ED visits resulted in hospitalization or death [19]. This may be due to the SNUBH-specific aspect that patients with cancer are treated via outpatient visits, and patients who were normally discharged might return to other inpatient care units or independent clinics after emergency treatments. If patients at risk could be identified during a previous scheduled session, many of these visits might be avoided. Moreover, it would be possible to perform elective procedures to manage patients at imminent risk of requiring ED visits. There have been several prior studies on predicting ED visits [20-22]; however, to our knowledge, there have been no studies predicting ED visits applicable to outpatient visits by patients with lung cancer. In addition, previous studies mainly focused on patient-reported outcomes, whereas this study mainly used clinical information, such as laboratory test results, as input variables for the prediction model. The clinical information–based machine learning model enhances the clinical use of the predictive model since patients with lung cancer continuously visit the hospital during their disease or after treatment for follow-up [23].

The identification of risk factors associated with ED visits among patients with lung cancer is another significant finding of this study. The risk factors were identified by analyzing the importance of features that influenced the decision-making of machine learning models, as shown in Figure 3, and the identified variables had a tendency to match theoretical knowledge found in the medical domain. Prior ED use was known to be a significant predictor of future ED visits for all types of patients with cancer [12]. Patients with cancer who had recently visited the ED more frequently may experience more severe symptoms, complications, or comorbidities requiring

immediate medical care, thereby increasing their likelihood of future ED visits. The number of elapsed days since the initial lung cancer diagnosis was inversely proportional to the likelihood of visiting the ED. Given that cancer-related treatments, such as surgery, chemotherapy, and radiotherapy, are typically administered within a year of diagnosis, acute diseases and transient side effects of the cancer-related treatment can result in emergency situations [24]. Similarly, the use of analgesics indicates that patients were suffering from severe pain and distress, which may have prompted visits to the ED for pain management. Several clinical pieces of information, such as laboratory test results routinely collected during cancer treatment and follow-up monitoring, were demonstrated to be useful predictors of ED visits in patients with lung cancer. Lower lymphocyte counts may indicate decreased immune function or increased vulnerability to infections, and higher leukocyte, monocyte, and neutrophil counts may be associated with inflammation; both of these abnormal values may lead to visits to the ED. Lower levels of hemoglobin, hematocrit, and erythrocyte counts may be signs of anemia, which may cause fatigue and result in visits to the ED. Significant weight loss and lower albumin levels may be associated with disease progression or malnutrition, increasing the risk of complications that may require ED visits. Similarly, a high CCI indicated the presence of multiple comorbid conditions. Elevated alkaline phosphatase levels may indicate liver or bone involvement, and decreased platelet counts may indicate thrombocytopenia, necessitating evaluation and treatment in the ED. While radiotherapy is a treatment for lung cancer, it can cause side effects and complications, such as radiation pneumonitis, that may necessitate visits to the ED. Thus, close monitoring of laboratory test results in patients with lung cancer receiving treatments or routine follow-up is necessary to prevent ED visits.

This study has several limitations. First, this study was conducted at a single institution in South Korea, a tertiary-level general hospital. Thus, regional bias may have resulted in reduced generalizability. Nonetheless, this study demonstrated the possibility of conducting observational cancer research with CDM data. Developing a predictive model using data from a single institution may lead to biased results based on regional or institution-specific characteristics, making it difficult to generalize; therefore, external validation is essential for clinical applications. Due to the heterogeneity of the format for managing medical records across institutions as well as privacy concerns, it is difficult for researchers to share data; even a collaborative study with other institutions requires immense effort, time, and resources. Ahmadi et al [25] reported that the OMOP CDM has the potential to facilitate international collaborative analyses, which is a crucial element of cancer precision medicine. Since our prediction model was developed using OMOP CDM data, it is expected that external validation will be possible for other institutions that have oncology CDM data through the observational health data sciences and informatics data network in the future. The data set used in this study is administered in accordance with the OMOP CDM, allowing different institutions to conduct collaborative research using a distributed research network.

Second, there were the inherent performance limitations imposed on the machine learning models. The primary outcome defined in this study was derived from all ED visits, only excluding the visits that resulted in patients being transferred to other institutions; therefore, there may have been non–cancer-related reasons for ED visits, such as fractures, injuries from traffic accidents, or wounds from animal attacks. In addition, a paucity of information may be one of the reasons for the results of unfavorable effects on the prediction model outcome. Staging or subtypes of lung cancer represent primary information for predicting the prognosis of patients; however, this information was not collected in our data source thus far. The acquisition of additional information could contribute to the enhancement of the predictive model's performance; therefore, the establishment of extended lung cancer–specific data sets in the CDM is needed for more precise prediction. Nevertheless, the identified risk factors for ED visits could be used in the future as an avenue for the proactive management of patients with lung cancer during treatment.

Lastly, research bias should be taken into account when interpreting the results. This study imputed missing values with the median value during the data processing phase, which could potentially have introduced bias into the analysis. However, the characteristics of clinical data are such that the lack of even a solitary test result frequently results in a substantial amount of missing data among other relevant data, indicating the presence of a clear-cut pattern. In such situations, it was noted that median imputation demonstrated performance levels that were comparable to those of more complex algorithms. Furthermore, it offered the advantage of being readily executable, thereby enhancing its efficiency in terms of time and resources. Additionally, it is critical to acknowledge that while removing outliers is a standard procedure for enhancing data quality, extreme values may have clinical ramifications due to the unique attributes of ED visits among patients diagnosed with lung cancer.

In summary, this study developed a machine learning model to predict ED visits, with a specific focus on patients with lung cancer, and identified influential factors that exerted a significant impact on the model output. Continuous monitoring of the identified influencing clinical features will allow health care providers to efficiently allocate resources, ensuring that patients with a high likelihood of requiring preventive care receive prompt attention. By identifying patients at risk, health care service providers can initiate targeted interventions, such as closer monitoring, treatment plan adjustments, and timely referrals to the most appropriate specialists. The predictive model for ED visits by patients with lung cancer developed in this study not only enhances patient outcomes but also optimizes resource utilization, reducing strain on the ED and minimizing the burden on the medical staff. Ultimately, this proactive approach contributes to preventing or mitigating emergency situations, resulting in an improved quality of life for patients with lung cancer and possibly reducing the need for hospitalization or more invasive interventions.

Acknowledgments

This research was supported by a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute funded by the Ministry of Health & Welfare, Republic of Korea (grant HI22C0471).

Authors' Contributions

AL and HP conceptualized the design of the study. SK and LS contributed to the data extraction. AL wrote the original draft. AL and HP performed the experiments and visualizations. AY and LS contributed to the data analysis and interpretation. SY revised the manuscript and supervised the research. All authors reviewed and approved the final version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
A list of selected features.
[PDF File (Adobe PDF File), 114 KB - medinform_v11i1e53058_app1.pdf]

Multimedia Appendix 2
Descriptive statistics of the selected features.
[DOCX File , 22 KB - medinform_v11i1e53058_app2.docx]

References

1. Sung H, Ferlay J, Siegel R, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2021 May;71(3):209-249 [FREE Full text] [doi: 10.3322/caac.21660] [Medline: 33538338]
2. Gallaway MS, Idaikkadar N, Tai E, Momin B, Rohan EA, Townsend J, et al. Emergency department visits among people with cancer: frequency, symptoms, and characteristics. J Am Coll Emerg Physicians Open 2021 Jun;2(3):e12438 [FREE Full text] [doi: 10.1002/emp2.12438] [Medline: 33969353]
3. Barta JA, Powell CA, Wisnivesky JP. Global epidemiology of lung cancer. Ann Glob Health 2019 Jan 22;85(1):8 [FREE Full text] [doi: 10.5334/aogh.2419] [Medline: 30741509]
4. Panattoni L, Fedorenko C, Greenwood-Hickman MA, Kreizenbeck K, Walker JR, Martins R, et al. Characterizing potentially preventable cancer- and chronic disease–related emergency department use in the year after treatment initiation: a regional study. J Oncol Pract 2018 Mar;14(3):e176-e185. [doi: 10.1200/jop.2017.028191]
5. Cronin KA, Lake AJ, Scott S, Sherman RL, Noone A, Howlader N, et al. Annual report to the nation on the status of cancer, part I: national cancer statistics. Cancer 2018 Jul 01;124(13):2785-2800 [FREE Full text] [doi: 10.1002/cncr.31551] [Medline: 29786848]
6. Walder JR, Faiz SA, Sandoval M. Lung cancer in the emergency department. Emerg Cancer Care 2023 Mar 06;2(1):2-12. [doi: 10.1186/s44201-023-00018-9]
7. Kotajima F, Kobayashi K, Sakaguchi H, Nemoto M. Lung cancer patients frequently visit the emergency room for cancer-related and -unrelated issues. Mol Clin Oncol 2014 Mar;2(2):322-326 [FREE Full text] [doi: 10.3892/mco.2014.241] [Medline: 24649355]
8. Shin SH, Lee H, Kang HK, Park JH. Twenty-eight-day mortality in lung cancer patients with metastasis who initiated mechanical ventilation in the emergency department. Sci Rep 2019 Mar 20;9(1):4941 [FREE Full text] [doi: 10.1038/s41598-019-39671-8] [Medline: 30894559]
9. Kim Y, Kang J, Kim M, Ryoo SM, Kang GH, Shin TG, et al. Development and validation of the VitaL CLASS score to predict mortality in stage IV solid cancer patients with septic shock in the emergency department: a multi-center, prospective cohort study. BMC Med 2020 Dec 14;18(1):390 [FREE Full text] [doi: 10.1186/s12916-020-01875-5] [Medline: 33308206]
10. Liu M, Bakow B, Hsu T, Chen J, Su K, Asiedu E, et al. Temporal trends in sepsis incidence and mortality in patients with cancer in the US population. Am J Crit Care 2021 Jul 01;30(4):e71-e79. [doi: 10.4037/ajcc2021632] [Medline: 34195781]
11. Alandonisi M, Al-Malki H, Bahaj W, Alghanmi H. Characteristics of emergency visits among lung cancer patients in comprehensive cancer center and impact of palliative referral. Cureus 2023 Apr;15(4):e37903 [FREE Full text] [doi: 10.7759/cureus.37903] [Medline: 37223145]
12. Hong JC, Niedzwiecki D, Palta M, Tenenbaum JD. Predicting emergency visits and hospital admissions during radiation and chemoradiation: an internally validated pretreatment machine learning algorithm. JCO Clinical Cancer Informatics 2018 Dec(2):1-11. [doi: 10.1200/cci.18.00037]
13. Sutradhar R, Rostami M, Barbera L. Patient-reported symptoms improve performance of risk prediction models for emergency department visits among patients with cancer: a population-wide study in Ontario using administrative data. J

Pain Symptom Manage 2019 Nov;58(5):745-755 [FREE Full text] [doi: 10.1016/j.jpainsymman.2019.07.007] [Medline: 31319103]
14. Sutradhar R, Barbera L. Comparing an artificial neural network to logistic regression for predicting ED visit risk among patients with cancer: a population-based cohort study. J Pain Symptom Manage 2020 Jul;60(1):1-9 [FREE Full text] [doi: 10.1016/j.jpainsymman.2020.02.010] [Medline: 32088358]
15. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc 2012;19(1):54-60 [FREE Full text] [doi: 10.1136/amiajnl-2011-000376] [Medline: 22037893]
16. ICD-10 version:2019. World Health Organization. URL: https://icd.who.int/browse10/2019/en#/C34.0 [accessed 2023-07-20]
17. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. 2017 Presented at: 31st Conference on Neural Information Processing Systems; Decemeber 4-9, 2017; Long Beach URL: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf
18. Roffman CE, Buchanan J, Allison GT. Charlson comorbidities index. J Physiother 2016 Jul;62(3):171 [FREE Full text] [doi: 10.1016/j.jphys.2016.05.008] [Medline: 27298055]
19. Lee SY, Ro YS, Shin SD, Moon S. Epidemiologic trends in cancer-related emergency department utilization in Korea from 2015 to 2019. Sci Rep 2021 Nov 09;11(1):21981 [FREE Full text] [doi: 10.1038/s41598-021-01571-1] [Medline: 34754058]
20. Katzan IL, Thompson N, Schuster A, Wisco D, Lapin B. Patient‐reported outcomes predict future emergency department visits and hospital admissions in patients with stroke. J Am Heart Assoc 2021 Mar 16;10(6):e018794. [doi: 10.1161/jaha.120.018794]
21. Chen R, Cheng K, Lin Y, Chang I, Tsai C. Predicting unscheduled emergency department return visits among older adults: population-based retrospective study. JMIR Med Inform 2021 Jul 28;9(7):e22491 [FREE Full text] [doi: 10.2196/22491] [Medline: 34319244]
22. Gao K, Pellerin G, Kaminsky L. Predicting 30-day emergency department revisits. Am J Manag Care 2018 Nov 01;24(11):e358-e364 [FREE Full text] [Medline: 30452204]
23. Owusuaa C, van der Padt-Pruijsten A, Drooger JC, Heijns JB, Dietvorst A, Janssens-van Vliet ECJ, et al. Development of a clinical prediction model for 1-year mortality in patients with advanced cancer. JAMA Netw Open 2022 Nov 01;5(11):e2244350 [FREE Full text] [doi: 10.1001/jamanetworkopen.2022.44350] [Medline: 36449290]
24. Bironzo P, Di Maio M. A review of guidelines for lung cancer. J Thorac Dis 2018 May;10(Suppl 13):S1556-S1563 [FREE Full text] [doi: 10.21037/jtd.2018.03.54] [Medline: 29951306]
25. Ahmadi N, Peng Y, Wolfien M, Zoch M, Sedlmayr M. OMOP CDM can facilitate data-driven studies for cancer prediction: a systematic review. Int J Mol Sci 2022 Oct 05;23(19):11834 [FREE Full text] [doi: 10.3390/ijms231911834] [Medline: 36233137]

Abbreviations

AUROC: area under the receiver operating characteristic curve
CCI: Charlson comorbidity index
CDM: common data model
ED: emergency department
LGBM: light gradient boosting machine
OMOP: observational medical outcomes partnership
PRAUC: area under the precision-recall curve
SHAP: Shapley additive explanations
SNUBH: Seoul National University Bundang Hospital

Edited by G Eysenbach, C Lovis; submitted 24.09.23; peer-reviewed by Z Su, D Hu, S Matsuda; comments to author 11.10.23; revised version received 31.10.23; accepted 24.11.23; published 06.12.23.

Please cite as:
Lee AR, Park H, Yoo A, Kim S, Sunwoo L, Yoo S
Risk Prediction of Emergency Department Visits in Patients With Lung Cancer Using Machine Learning: Retrospective Observational Study
JMIR Med Inform 2023;11:e53058
URL: https://medinform.jmir.org/2023/1/e53058
doi:10.2196/53058
PMID:38055320

Corrigenda and Addenda

Correction: Social Media Monitoring of the COVID-19 Pandemic and Influenza Epidemic With Adaptation for Informal Language in Arabic Twitter Data: Qualitative Study

Lama Alsudias[1,2], BSc, MSc, PhD; Paul Rayson[2], BSc, PhD

[1]Information Technology Department, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia

[2]School of Computing and Communications, Lancaster University, Lancaster, United Kingdom

Corresponding Author:
Lama Alsudias, BSc, MSc, PhD
Information Technology Department
College of Computer and Information Sciences
King Saud University
Prince Turki Bin Abdulaziz Al Awwal Road
Riyadh, 12371
Saudi Arabia
Phone: 966 118051044
Email: lalsudias@ksu.edu.sa

Related Article:

Correction of: https://medinform.jmir.org/2021/9/e27670

(JMIR Med Inform 2023;11:e45742) doi:10.2196/45742

In "Social Media Monitoring of the COVID-19 Pandemic and Influenza Epidemic With Adaptation for Informal Language in Arabic Twitter Data: Qualitative Study" (JMIR Med Inform 2021;9(9):e27670) the authors made one addition to the Acknowledgments section and changes to the Corresponding Author address and degree list.

The Acknowledgments has been changed from:

> *The authors thank Nouran Khallaf, who is a PhD student at Leeds University (mlnak@leeds.ac.uk), for her help in labeling the tweets. LA wishes to thank King Saud University for funding her PhD study.*

To:

> *The authors thank Nouran Khallaf, who is a PhD student at Leeds University (mlnak@leeds.ac.uk), for her help in labeling the tweets. This research project was supported by a grant from the "Research Center of College of Computer and Information Sciences", Deanship of Scientific Research, King Saud University.*

Additionally, the Corresponding Author Address and degree list has been changed from:

> *Lama Alsudias, BSc, MSc*
> *School of Computing and Communications*
> *Lancaster University*
> *InfoLab21*
> *Lancaster*
> *GB*
> *Phone: 44 1524 510357*
> *Email:l.alsudias@lancaster.ac.uk*

To:

> *Lama Alsudias, BSc, MSc, PhD*
> *Information Technology Department*
> *College of Computer and Information Sciences*
> *King Saud University*
> *Prince Turki Bin Abdulaziz Al Awwal Road*
> *Riyadh, 12371*
> *Saudi Arabia*
> *Phone:: 966 118051044*
> *Email: lalsudias@ksu.edu.sa*

The correction will appear in the online version of the paper on the JMIR Publications website on February 3, 2023 together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article

Submitted 15.01.23; this is a non–peer-reviewed article;accepted 16.01.23; published 03.02.23.

Please cite as:
Alsudias L, Rayson P
Correction: Social Media Monitoring of the COVID-19 Pandemic and Influenza Epidemic With Adaptation for Informal Language in Arabic Twitter Data: Qualitative Study
JMIR Med Inform 2023;11:e45742
URL: https://medinform.jmir.org/2023/1/e45742
doi:10.2196/45742
PMID:36735930

Corrigenda and Addenda

Correction: Data Analysis of Physician Competence Research Trend: Social Network Analysis and Topic Modeling Approach

So Jung Yune[1*], PhD; Youngjon Kim[2*], PhD; Jea Woog Lee[3*], PhD

[1]Department of Medical Education, Pusan National University, Busan, Republic of Korea

[2]Department of Medical Education, Wonkwang University School of Medicine, Iksan, Republic of Korea

[3]Intelligence Informatics Processing Lab, Chung-Ang University, Seoul, Republic of Korea

[*]all authors contributed equally

Corresponding Author:
Jea Woog Lee, PhD
Intelligence Informatics Processing Lab
Chung-Ang University
84, Heukseok-ro, Dongjak-gu
Seoul, 06974
Republic of Korea
Phone: 82 10 5426 7318
Email: yyizeuks@cau.ac.kr

Related Article:

Correction of: https://medinform.jmir.org/2023/1/e47934

(JMIR Med Inform 2023;11:e53484) doi:10.2196/53484

In "Data Analysis of Physician Competence Research Trend: Social Network Analysis and Topic Modeling Approach" (JMIR Serious Games 2023;11(1):e47934) the authors made one correction:

In the corrected article, a new Acknowledgments section appeared as follows:

This paper was supported by Wonkwang University in 2022.

The correction will appear in the online version of the paper on the JMIR Publications website on November 1, 2023, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Submitted 08.10.23; this is a non–peer-reviewed article;accepted 16.10.23; published 31.10.23.

Please cite as:
Yune SJ, Kim Y, Lee JW
Correction: Data Analysis of Physician Competence Research Trend: Social Network Analysis and Topic Modeling Approach
JMIR Med Inform 2023;11:e53484
URL: https://medinform.jmir.org/2023/1/e53484
doi:10.2196/53484
PMID:

Editorial

"To Err Is Evolution": We Need the Implementation Report to Learn

Caroline Perrin Franck[1,2], PhD; Antoine Geissbuhler[1,2,3], MD; Christian Lovis[1,3], MD

[1]Campus Biotech, University of Geneva, Geneva, Switzerland

[2]Geneva Digital Health Hub, Geneva, Switzerland

[3]Geneva University Hospitals, Geneva, Switzerland

Corresponding Author:
Caroline Perrin Franck, PhD
Campus Biotech
University of Geneva
Chemin des Mines 9
Geneva, 1202
Switzerland
Phone: 41 787997725
Email: caroline.perrin@unige.ch

Abstract

JMIR Medical Informatics is pleased to offer *implementation reports* as a new article type. Implementation reports present real-world accounts of the implementation of health technologies and clinical interventions. This new article type is intended to promote the rapid documentation and dissemination of the perspectives and experiences of those involved in implementing digital health interventions and assessing the effectiveness of digital health projects.

(JMIR Med Inform 2023;11:e47695) doi:10.2196/47695

KEYWORDS

implementation science; knowledge management; knowledge sharing; digital health; implementation report

Introduction

The accelerating adoption of digital health, combined with evolving terminology and differing definitions, has created a paradox: there is an exponential increase in knowledge and information, but finding relevant data on implementation processes, and in particular errors or failures, remains a major challenge. This lack of effective documentation of implementation knowledge makes it difficult to reliably study and understand global trends in digital health project failures, exacerbated by a bias toward publishing mostly positive studies [1]. As a result, similar and avoidable mistakes are often repeated.

Digitalization offers numerous opportunities to improve the efficiency and equity of health systems. Yet, many digital health implementations stagnate in the pilot phase or fail to sustain or demonstrate impact. This not only wastes valuable resources but also further fragments already complex health systems, potentially leading to adverse health outcomes [2].

Recurrent errors contribute to inefficient implementation and scale-up, and it is crucial to consider that while new technologies offer immense potential benefits, they can also introduce risks or unintended consequences that have a direct impact on patient outcomes. For example, child mortality increased significantly at one site following the implementation of a commercial computerized physician order entry (CPOE) system [3]. This increase in mortality was primarily due to delayed administration of critical medications [4], amplified by a policy change shortly before implementation, suggesting that policy changes should be avoided during or in close proximity to a CPOE implementation process [5].

However, unintended consequences can also be positive, as demonstrated by the implementation of a telemedicine service in rural Nepal [6]. The service attracted a higher proportion of female patients, possibly due to cultural factors or minimal disruption to their daily lives, which suggests that telemedicine may improve access to health care for female patients [6].

Learning from past implementations is critical. Particularly from an ethical and human rights perspective, as in the absence of established implementation norms and best practices, implementers need to define their own standards for responsible and effective digital health implementations.

The Potential Transformative Impact of Implementation Reports

Sharing and connecting fragmented knowledge across institutional boundaries can revolutionize an industry. This is

evidenced by the aviation sector, a safety-critical industry like health care, where such collaboration has led to transformative results [7]:

> *Back in the 1930s, flying was really dangerous and passengers were scared away by the many accidents. Flight authorities across the world had understood the potential of commercial passenger air traffic, but they also realized flying had to become safer before most people would dare to try it. In 1944 they all met in Chicago to agree on common rules and signed a contract with a very important Annex 13: a common form for incidents reports, which they agreed to share, so they could all learn from each other's mistakes. Since then, every crash or incident involving a commercial passenger airplane has been investigated and reported; risk factors have been systematically identified; and improved safety procedures have been adopted worldwide.*

The aviation industry has set an exemplary precedent for how sharing mistakes can improve safety and build trust. However, much like the aviation industry in the 1930s, the digital health sector is still in the early stages: the potential to strengthen health care systems is clear, but the bigger picture and potential implications and safety risks are not yet fully understood.

JMIR Medical Informatics has created a new article type—the implementation report—to address the challenges associated with managing and effectively sharing implementation knowledge. We need implementation reports to promote greater transparency and accountability; to improve identification of best practices; to optimize resource allocation; and to help gain the trust of patients, practitioners, and other stakeholders. Implementation reports provide a framework for systematically documenting and sharing implementation knowledge, including errors and failures, and present real-world accounts of the implementation of health technologies and clinical interventions. This new article type aims to promote the rapid documentation and dissemination of the perspectives and experiences of those involved in implementing digital health interventions and assessing the effectiveness of digital health projects.

If trial and error is the key engine of evolution, achieving progress requires mechanisms to store and transmit information. DNA is the main information substrate of the evolving living world. We need similar tools, such as the implementation report, to learn and accumulate knowledge from successes and failures in digital health. By embracing the concept that "to err is evolution," we can collectively harness the power of our experiences to drive innovation and improve health outcomes.

Conflicts of Interest

None declared.

References

1. Dendere R, Janda M, Sullivan C. Are we doing it right? We need to evaluate the current approaches for implementation of digital health systems. Aust Health Rev 2021 Dec;45(6):778-781. [doi: 10.1071/AH20289] [Medline: 34488938]
2. Thompson M. The environmentally impacts of digital health. Digit Health 2021 Aug 10;7:20552076211033421 [FREE Full text] [doi: 10.1177/20552076211033421] [Medline: 34408902]
3. Han YY, Carcillo JA, Venkataraman ST, Clark RSB, Watson RS, Nguyen TC, et al. Unexpected increased mortality after implementation of a commercially sold computerized physician order entry system. Pediatrics 2005 Dec;116(6):1506-1512. [doi: 10.1542/peds.2005-1287] [Medline: 16322178]
4. Beeler P, Bates D, Hug B. Clinical decision support systems. Swiss Med Wkly 2014 Dec 23;144:w14073 [FREE Full text] [doi: 10.4414/smw.2014.14073] [Medline: 25668157]
5. Sittig DF, Ash JS, Zhang J, Osheroff JA, Shabot MM. Lessons from "Unexpected increased mortality after implementation of a commercially sold computerized physician order entry system". Pediatrics 2006 Aug;118(2):797-801. [doi: 10.1542/peds.2005-3132] [Medline: 16882838]
6. Gupta P, Uranw S, Gupta S, Das R, Bhattarai A, Bhatta N, et al. Study of the impact of a telemedicine service in improving pre-hospital care and referrals to a tertiary care university hospital in Nepal. J Family Med Prim Care 2021 Dec;10(12):4531-4535 [FREE Full text] [doi: 10.4103/jfmpc.jfmpc_9_21] [Medline: 35280611]
7. Rosling H. Factfulness: Ten Reasons We're Wrong about the World—and Why Things Are Better than You Think. New York, NY: Flatiron Books; 2018.

Abbreviations

CPOE: computerized physician order entry

Edited by T Leung; submitted 29.03.23; this is a non–peer-reviewed article;accepted 30.03.23; published 04.04.23.

Please cite as:
Perrin Franck C, Geissbuhler A, Lovis C
"To Err Is Evolution": We Need the Implementation Report to Learn
JMIR Med Inform 2023;11:e47695
URL: https://medinform.jmir.org/2023/1/e47695
doi:10.2196/47695
PMID:37014675

Editorial

Introducing the "AI Language Models in Health Care" Section: Actionable Strategies for Targeted and Wide-Scale Deployment

Alexandre Castonguay[1*], PhD; Christian Lovis[2*], MPH, MD

[1]Faculté des sciences infirmières, Université de Montréal, Montréal, QC, Canada

[2]Faculty of Medicine, University of Geneva, Geneva, Switzerland

[*]all authors contributed equally

Corresponding Author:
Alexandre Castonguay, PhD
Faculté des sciences infirmières
Université de Montréal
2375, chemin de la Côte-Sainte-Catherine
Montréal, QC, H3T1A8
Canada
Email: alexandre.castonguay.2@umontreal.ca

Abstract

The realm of health care is on the cusp of a significant technological leap, courtesy of the advancements in artificial intelligence (AI) language models, but ensuring the ethical design, deployment, and use of these technologies is imperative to truly realize their potential in improving health care delivery and promoting human well-being and safety. Indeed, these models have demonstrated remarkable prowess in generating humanlike text, evidenced by a growing body of research and real-world applications. This capability paves the way for enhanced patient engagement, clinical decision support, and a plethora of other applications that were once considered beyond reach. However, the journey from potential to real-world application is laden with challenges ranging from ensuring reliability and transparency to navigating a complex regulatory landscape. There is still a need for comprehensive evaluation and rigorous validation to ensure that these models are reliable, transparent, and ethically sound. This editorial introduces the new section, titled "AI Language Models in Health Care." This section seeks to create a platform for academics, practitioners, and innovators to share their insights, research findings, and real-world applications of AI language models in health care. The aim is to foster a community that is not only excited about the possibilities but also critically engaged with the ethical, practical, and regulatory challenges that lie ahead.

(JMIR Med Inform 2023;11:e53785) doi:10.2196/53785

KEYWORDS

generative AI; health care digitalization; AI in health care; digital health standards; AI implementation; artificial intelligence

The Promise and Potential of AI in Health Care

In this editorial, we introduce a new section in *JMIR Medical Informatics*, "AI Language Models in Health Care," dedicated to exploring the transformative potential of generative artificial intelligence (AI) and the developments of cognitive AI in the health care industry. At its current stage of development, AI has already shown the potential to guide human interactions, facilitate understanding, and foster respectful communication. Most of all, with the phenomenon surrounding the global release of ChatGPT [1], it is the first time that information technology has demonstrated the ability to leverage the informational content of texts and human narratives in such a substantial way. In a world where over 65% of the population is connected to the internet [2], this potential is substantial and wide-reaching. Within the health care sector specifically, AI and large language models in particular have progressed toward fulfilling their ambitious promises. They have delivered concrete outcomes such as improving decision-making processes at all levels of influence including for patients, clinicians, health systems, and society as a whole [3]. Yet, the journey of generative AI is still nascent. While the rapid adoption of generative AI indicates its potential, concerns regarding its accuracy and preparedness in managing its associated risks remain [4]. Additionally, being in its early stages of deployment, the long-term positive impacts are anticipated but not yet fully realized or evidenced.

The Aim of This New Section

In this new journal section, we focus on generative models and how they can redefine the boundaries of health care, all while acknowledging the challenges and limitations. By spotlighting

generative models, we hope to illuminate their potential while realistically addressing the challenges associated with their implementation in diverse care settings.

The objective of this section is to highlight groundbreaking innovations empowered by AI language models with a keen focus on ensuring their safe, effective, and enduring implementation. We seek to delve into solid actionable strategies for implementation at various scales of deployment and sustainability, striving to facilitate the swift yet secure integration of these technologies into health care environments. Our ambitions extend beyond theoretical contributions; we seek mature, ready-to-implement initiatives that not only break new ground but also offer clear actionable strategies for wide-scale deployment. The goal is to leapfrog over efforts that should have been made 20 years ago and to streamline the integration process of innovations that will propel the health care sector 20 years into the future.

This section is dedicated to the implementation of AI language models in health care and focuses on prioritizing efforts that highlight the best and most accessible opportunities. Its goal is to leverage emerging technologies while aiming to deliver substantial benefits to as many people as possible.

By featuring papers with clear detailed methodologies, we aim to inspire and facilitate the adaptation and replication of these innovations across diverse health care environments. In doing so, we hope that this new section serves not only as a road map but also as an inspiration, catalyzing the ambition to redefine the limits of health care through the responsible strategic use of AI. More information on suggested topics for submission can be found in the corresponding section of this editorial.

The Nuanced Role of Generative AI in Health Care

While digitalization remains the foundation upon which our current health care advancements are built [5], it is essential to differentiate between the overarching digital transformation and the nuanced capabilities of generative AI. Generative AI models, a subset of the broader AI landscape, hold unique promise and challenges for the health care sector. Unlike traditional AI systems that operate within the confines of given data sets, generative AI can create new, unseen content [6], enhancing clinical decision-making, patient interactions, and even medical research [3]. This ability to generate information places a heightened responsibility on ensuring the reliability, accuracy, and ethics of its outputs. As we navigate the vast world of health care AI, it is imperative to hone our focus on these generative models and their distinct potential to reshape patient care, medical education, and health system operations. By emphasizing their role, we aim to highlight not only their innovative capabilities but also the accompanying challenges that need to be addressed for a meaningful and safe integration into health care environments.

Current State of Affairs: The Great Paradox

Despite significant advancements in AI often leaving populations worldwide in awe, many areas within the health care sector still notoriously fail to impress, particularly in terms of digital health standards, equitable access, and system interoperability [7]. Some medical sectors in specific places are at the forefront of scientific and technological advances, but most countries or fields like long-term care lag, trapped in the digital health standards of the 1970s. The major challenges are not the use of faxes or digital tools but rather the equity of access for society and interoperability for the most advanced systems.

As a sector receiving substantial investment, being the second-largest spending category among Organisation for Economic Co-operation and Development countries, the health care industry is more than a line item in a budget [8]. Fundamentally, it is charged with the critical task of promoting and safeguarding health—one of the indispensable pillars supporting societal stability on which all stands. Without it, everything else teeters on the edge of collapse.

The guiding principle of health care has long been “Primum non noccere,” which translates to “First, do no harm.” Yet in today’s rapidly evolving digital landscape, the traditional interpretation of this maxim can no longer serve as an excuse for inaction. In our current state, maintaining the status quo or doing nothing is, paradoxically, causing harm.

The Call to Action: Leading the Digital Charge

Digitalization, with its foundational role in ensuring data interoperability, has enabled the rise of AI in health. In turn, AI has provided a platform for the advent of advanced generative models. Today, amid this AI revolution, it is striking and unacceptable that, despite its crucial role and significant investment, health care often finds itself trailing [9,10]. As the technological landscape evolves at an unprecedented pace, we should not merely object to this essential industry remaining technologically outdated; we should insist that it leads the charge, setting the pace rather than lagging. However, this is not just about technological advancements; it is about the essence of care.

The potential of generative models in health care is vast, and our discussion should reflect an informed enthusiasm. Recognizing the capabilities and promise they bring does not mean we are overlooking the complexities. It means we are optimistic about what can be achieved without being naive about the challenges. While we push for innovation in health care, our insistence is on not only digital advancements but also meaningful, responsible, and nuanced progression. Now more than ever, advancing with clarity and purpose is the only way health care can genuinely uphold “Primum non noccere” in this digital age.

The Present Challenges: Strategies for Effective Implementation

A wealth of digital health solutions already exists [11], each seemingly more promising than the last. However, as technology advances, understanding these solutions requires increasingly specialized skills.

For instance, when users turn to chatbots such as ChatGPT for specific health care queries—be it programming a medical device or interpreting a health symptom—the ideal response from such systems should be precise, accurate, and aware of its limitations. However, instead of straightforward confessions like, "I'm unfamiliar with that programming language" or "That symptom is beyond my knowledge," these models might sometimes fabricate answers to feign understanding [12]. This tendency to generate unverified responses, without a built-in "truth threshold," not only risks misinformation but also erodes trust in these technologies in a field where clarity is paramount.

This situation underscores the imperative for the industry to enhance the reliability and transparency of such tools and, equally, for users to be aptly equipped to use and interpret these advanced technologies. They must act as a "truth threshold" for themselves, discerning the reliability of the information provided.

This represents an indispensable uphill battle to evaluate and prioritize solutions for large-scale implementation since the need has never been greater for efficient, safe, and intelligent solutions that improve performance and quality of life for patients and health care professionals while benefiting institutions and the health care system as a whole.

The Journey Forward: Meticulous Implementation and Sustainability

In light of AI's transformative potential—and recognizing it both as a beacon of groundbreaking advancements and, if unchecked, as one of the most significant potential threats to humanity—our call to action underscores the great importance of not only progressing AI's technical development and governance. It is, for example, equally crucial to address the domains of regulatory constraints, interoperability, accountability, and liability with utmost diligence [13]. Drawing a parallel from the aeronautics industry, where manufacturers, air traffic control, and stringent flight regulations collectively ensure some of the safest travel experiences [14], we believe that the journey of AI integration in health care must adopt a similar, if not greater, level of care and rigor.

However, it is pivotal to highlight that sheer development and regulations will not suffice. The essence of successful innovation lies in its effective implementation. This includes rigorous preparation, meticulous execution, and sustained support to ensure both immediate success and long-term sustainability. Given the health care industry's paramount role in life preservation and its significant costs, it is imperative that it not only aligns with technological advancements but also leads them. Leading does not mean merely developing new solutions but also ensuring they are seamlessly integrated and enduringly effective. The health care sector should be at the helm of innovation, shaping the future and guaranteeing that the innovations are not only introduced but also enduringly implemented for the benefit of all.

Suggested Topics for Submission

This section provides a list of example topics tailored to align with our broad focus on the successful implementation of AI language models in diverse health care settings. We encourage contributors to delve into the process, use, outcomes, and influential factors of AI integration, aiming for a holistic and practical discourse. The topics listed are not exhaustive, and we welcome diverse perspectives and innovative approaches.

- Comparative analyses of policies and regulations: Explore the varying landscapes of AI regulation in health care across different countries. Delve into how specific policies either facilitate or hinder the implementation of AI language models, providing insights into creating conducive environments for global adoption.
- Strategies for reliability, transparency, and ethical use: Discuss and develop strategies to ensure the ethical, transparent, and reliable application of AI language models in health care. Share best practices, guidelines, and frameworks that help in establishing trust among stakeholders and ensuring patient safety.
- Comprehensive methodologies for implementation: Provide detailed blueprints and methodologies for the successful integration of AI language models in health care settings. Focus on the entire life cycle, from initial planning and integration to long-term management, ensuring applicability for researchers, industry professionals, policy makers, and decision makers.
- Case studies on real-world impact and effectiveness: Share real-world examples illustrating the impact of AI language models on health care delivery. Highlight both successes and challenges, discussing the tangible outcomes, lessons learned, and strategies for overcoming obstacles.
- Evaluations of outcomes and impact on health care delivery: Conduct thorough evaluations of AI language model implementations, examining a wide range of outcomes including acceptability, adoption, cost-effectiveness, and impact on health care efficiency and patient satisfaction. Explore how these technologies contribute to or challenge the equity, timeliness, and patient-centeredness of health care services.

By addressing these topics, authors contribute to a collaborative, rigorous, and transformative discourse, propelling the health care sector toward seamless integration of AI language models and ensuring a lasting positive impact on global health care delivery.

Authors' Contributions

Both authors participated equally in conceptualization. While AC contributed more significantly to writing the original draft, CL contributed more to reviewing and editing.

Conflicts of Interest

CL is the editor-in-chief of *JMIR Medical Informatics*. AC is an editorial board member of *JMIR Medical Informatics*.

References

1. Hu K. ChatGPT sets record for fastest-growing user base - analyst note. Reuters. 2023 Feb 02. URL: https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/ [accessed 2023-12-13]
2. Population of global offline continues steady decline to 2.6 billion people in 2023. ITU. 2023 Sep 12. URL: https://www.itu.int/en/mediacentre/Pages/PR-2023-09-12-universal-and-meaningful-connectivity-by-2030.aspx [accessed 2023-12-13]
3. Thirunavukarasu A, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023 Aug;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
4. Chui M, Yee L, Hall B, Singla A, Sukharevsky A. The state of AI in 2023: generative AI's breakout year. McKinsey & Company. 2023. URL: http://tinyurl.com/2jff7965 [accessed 2023-12-13]
5. WHO launches a new Global Initiative on Digital Health supported by the G20 Presidency. World Health Organization. 2023 Aug 19. URL: http://tinyurl.com/bdewspe6 [accessed 2023-12-13]
6. Russell S, Norvig P. Artificial Intelligence: A Modern Approach, 4th edition. Upper Saddle River, NJ: Pearson; Apr 28, 2020.
7. Paik K, Hicklen R, Kaggwa F, Puyat C, Nakayama L, Ong B, et al. Digital determinants of health: health data poverty amplifies existing health disparities-a scoping review. PLOS Digit Health 2023 Oct;2(10):e0000313 [FREE Full text] [doi: 10.1371/journal.pdig.0000313] [Medline: 37824445]
8. General government spending. OECD Data. 2023. URL: https://data.oecd.org/gga/general-government-spending.htm [accessed 2023-11-01]
9. Cam A, Chui M, Hall B. Global AI survey: AI proves its worth, but few scale impact. McKinsey & Company. 2019 Nov 22. URL: http://tinyurl.com/59yjexdw [accessed 2023-12-12]
10. Goldfarb A, Teodoridis F. Why is AI adoption in health care lagging? Brookings. 2022 Mar 09. URL: https://www.brookings.edu/articles/why-is-ai-adoption-in-health-care-lagging/ [accessed 2023-12-12]
11. Classification of digital interventions, services and applications in health: a shared language to describe the uses of digital technology for health, 2nd ed. World Health Organization. 2023 Oct 24. URL: https://www.who.int/publications/i/item/9789240081949 [accessed 2023-12-13]
12. Alkaissi H, McFarlane SI. Artificial hallucinations in ChatGPT: implications in scientific writing. Cureus 2023 Feb 19;15(2):1-4 [FREE Full text] [doi: 10.7759/cureus.35179] [Medline: 36811129]
13. Working Group on Digital and AI in Health. Reimagining global health through artificial intelligence: the roadmap to AI maturity. Broadband Commission. 2020 Sep. URL: https://broadbandcommission.org/wp-content/uploads/2021/02/WGAIinHealth_Report2020.pdf [accessed 2023-12-12]
14. Global Aviation Safety Plan 2023–2025. International Civil Aviation Organization. 2023. URL: https://www.icao.int/safety/GASP/Documents/10004_en.pdf [accessed 2023-12-13]

Abbreviations

AI: artificial intelligence

Edited by T Leung; submitted 18.10.23; this is a non–peer-reviewed article;accepted 02.12.23; published 21.12.23.

Please cite as:
Castonguay A, Lovis C
Introducing the "AI Language Models in Health Care" Section: Actionable Strategies for Targeted and Wide-Scale Deployment
JMIR Med Inform 2023;11:e53785
URL: https://medinform.jmir.org/2023/1/e53785
doi:10.2196/53785
PMID:38127431

Comparing Natural Language Processing and Structured Medical Data to Develop a Computable Phenotype for Patients Hospitalized Due to COVID-19: Retrospective Analysis

Feier Chang[1,*], BA, MB; Jay Krishnan[2,*], MD; Jillian H Hurst[3], PhD; Michael E Yarrington[2], MD; Deverick J Anderson[2], MD; Emily C O'Brien[4,5], PhD; Benjamin A Goldstein[1,3,4,5], PhD

1

2

3

4

5

*these authors contributed equally

Corresponding Author:
Benjamin A Goldstein, PhD

Abstract

Background: Throughout the COVID-19 pandemic, many hospitals conducted routine testing of hospitalized patients for SARS-CoV-2 infection upon admission. Some of these patients are admitted for reasons unrelated to COVID-19 and incidentally test positive for the virus. Because COVID-19–related hospitalizations have become a critical public health indicator, it is important to identify patients who are hospitalized because of COVID-19 as opposed to those who are admitted for other indications.

Objective: We compared the performance of different computable phenotype definitions for COVID-19 hospitalizations that use different types of data from electronic health records (EHRs), including structured EHR data elements, clinical notes, or a combination of both data types.

Methods: We conducted a retrospective data analysis, using clinician chart review–based validation at a large academic medical center. We reviewed and analyzed the charts of 586 hospitalized individuals who tested positive for SARS-CoV-2 in January 2022. We used LASSO (least absolute shrinkage and selection operator) regression and random forests to fit classification algorithms that incorporated structured EHR data elements, clinical notes, or a combination of structured data and clinical notes. We used natural language processing to incorporate data from clinical notes. The performance of each model was evaluated based on the area under the receiver operator characteristic curve (AUROC) and an associated decision rule based on sensitivity and positive predictive value. We also identified top words and clinical indicators of COVID-19–specific hospitalization and assessed the impact of different phenotyping strategies on estimated hospital outcome metrics.

Results: Based on a chart review, 38.2% (224/586) of patients were determined to have been hospitalized for reasons other than COVID-19, despite having tested positive for SARS-CoV-2. A computable phenotype that used clinical notes had significantly better discrimination than one that used structured EHR data elements (AUROC: 0.894 vs 0.841; $P<.001$) and performed similarly to a model that combined clinical notes with structured data elements (AUROC: 0.894 vs 0.893; $P=.91$). Assessments of hospital outcome metrics significantly differed based on whether the population included all hospitalized patients who tested positive for SARS-CoV-2 or those who were determined to have been hospitalized due to COVID-19.

Conclusions: These findings highlight the importance of cause-specific phenotyping for COVID-19 hospitalizations. More generally, this work demonstrates the utility of natural language processing approaches for deriving information related to patient hospitalizations in cases where there may be multiple conditions that could serve as the primary indication for hospitalization.

(JMIR Med Inform 2023;11:e46267) doi:10.2196/46267

KEYWORDS

natural language processing; NLP; computable phenotype; machine learning; COVID; coronavirus; hospitalize; hospitalization; electronic health record; EHR; health record; structured data; data element; free text; unstructured data; provider note; classify; classification; algorithm; COVID-19

Introduction

Hospitalization due to COVID-19 has become a key public health indicator. One of the primary goals of vaccination against SARS-CoV-2, the etiological agent of COVID-19, is to reduce the incidence of severe disease and death, with hospitalization serving as a primary end point in vaccine efficacy trials [1]. Further, hospitalization has become a primary indicator of community transmission levels of SARS-CoV-2 infection [2], including disease severity and health system capacity [3-6]. Similarly, hospitalization due to COVID-19 is a typical outcome of interest in public health studies of COVID-19 using real-world data sources, such as electronic health record (EHR) data [7-10]. Finally, because of the rise of rapid, at-home testing for SARS-CoV-2 infection, COVID-19 cases that do not rise to the level of requiring medical attention are likely to be missed or underreported, affecting assessments of COVID-19 prevalence [11]. Thus, there is a critical need to rapidly and accurately identify hospitalizations due to COVID-19.

Due to concerns related to the hospital-based spread of SARS-CoV-2, many institutions routinely perform SARS-CoV-2 testing in patients who are admitted to the hospital, regardless of the primary reason for admission [12,13]. Although SARS-CoV-2 testing is important for guiding care and ensuring that health care professionals take precautions to prevent infection, such routine testing potentially complicates retrospective studies using real-world data sources. Specifically, it becomes challenging to distinguish a patient who was admitted because of COVID-19 from a patient who incidentally tested positive for SARS-CoV-2 infection. In both cases, patients would have a positive laboratory test result and would (presumably) have an *International Classification of Diseases, 10th Revision* (*ICD-10*) code for COVID-19. Previous reports have noted that incidental positives may account for around 26% of all COVID-19–positive patients [14].

Given the public health importance of identifying hospitalizations due to COVID-19 rather than hospitalizations in which SARS-CoV-2 infection was identified incidentally, methods (ie, computable phenotypes) are needed to distinguish the two conditions in retrospective data sources. Such phenotypes would be instrumental in retrospective studies of patients with COVID-19 and in public health surveillance. In this study, we seek to (1) motivate the need to identify patients who were admitted because of COVID-19 versus patients who incidentally tested positive for SARS-CoV-2 during admission, (2) explore the potential of using both structured data (ie, diagnosis codes, medications, and procedure codes) and unstructured data (ie, clinical notes) to construct computable phenotypes, and (3) illustrate the inferential biases that may arise if phenotyping methods cannot distinguish the reason for hospitalization.

Methods

Study Setting

We performed a retrospective study of patients aged >18 years who were hospitalized with a documented positive SARS-CoV-2 test result during January 2022. We conducted our study at Duke University Health System (DUHS), which consists of 1 quaternary academic medical center and 2 associated community-based hospitals.

Ethical Considerations

This study was designated as exempt human subjects research by the DUHS Institutional Review Board (IRB number: Pro00109397).

Study Data

Source Data

Using DUHS EHR data, we identified all patients who were admitted during the week of January 16 to 22, 2022, with documentation of a positive SARS-CoV-2 test result in the prior 20 days. Charts from this week were specifically reviewed in part due to a data request from the North Carolina Division of Public Health to understand the epidemiology of COVID-19–related hospitalizations. We excluded individuals with a resolved COVID-19 isolation status, as well as those who were admitted prior to January 1, 2022, to create a cohort of patients who were likely infected with the Omicron variant of SARS-CoV-2. During this period, the Omicron variant was the predominant SARS-CoV-2 variant in circulation within the United States and was associated with the largest wave [8] of SARS-CoV-2 infections to date. For each patient, we extracted the following data: medical record number, date of admission, hospital unit, and level of care.

To generate a criterion standard for classification, 6 trained health care professionals manually reviewed patient records for the index admission to adjudicate whether SARS-CoV-2 infection was the primary reason for admission or an incidental finding. Health care professionals attributed hospitalizations as those due to COVID-19 if admissions were due to primary manifestations of SARS-CoV-2 infection, such as hypoxia or the need for supplemental oxygen, or due to COVID-19–associated complications, such as dehydration or weakness.

Analytic Data

For each admission reviewed, we extracted structured EHR data elements recorded during hospitalization and captured within the Duke Clinical Research Datamart—an EHR database that is based on an extension of the PCORnet Common Data Model (National Patient-Centered Clinical Research Network) [15]. Clinical notes were extracted from the Duke University Electronic Data Warehouse. We extracted admission data, daily progress data, and discharge summary notes. Extracted structured data elements included demographics, service encounter characteristics, diagnoses, laboratory tests, COVID-19 vaccination status, and medications (Table S1 in Multimedia Appendix 1). Clinical notes included emergency department admission notes, progress notes, operative notes, history and physical examination notes, and discharge summaries.

Clinical Note Analysis

To analyze the clinical notes, we used the term frequency–inverse document frequency (TF-IDF) approach. The TF-IDF approach [16] generates, across the set of notes for each patient, a numeric value for each word. The word value is

based on how common the word is in a patient's set of notes (term frequency), divided by how common the word is across all of the patient's notes (inverse document frequency), resulting in a numeric representation for each word on a per-patient basis. Although this is a simple word-based representation, this approach has the following two advantages over deep learning embedding–based approaches: (1) it is possible to directly assess the importance of individual words, and (2) the TF-IDF tends to be more robust with small data sets. Notes were extracted as CSV files and concatenated for the entire encounter. We used the *nltk* package in Python (Python Software Foundation) [17] to tokenize words into a dictionary. For each document, we calculated word counts and removed any words that appeared fewer than 50 times. We then generated the corresponding weight matrix, which served as a numeric input for downstream analyses.

Analytic Approach

We first described the clinical characteristics of patients hospitalized due to COVID-19 versus those with incidental COVID-19 by using standardized mean differences (SMDs), with an SMD of 0.10 indicating a clinically meaningful difference. Next, we developed 3 classification models for COVID-19–specific hospitalization; one was based entirely on structured EHR data elements, a second was based on clinical notes alone, and a third used both structured data elements and clinical notes. We used LASSO (least absolute shrinkage and selection operator) [18] logistic regression and random forests [19] to estimate the models. Due to the relatively small sample size, we presented our results based on 10-fold cross-validation. We performed the TF-IDF approach separately within each cross-validation fold.

We evaluated the six classification models by calculating the area under the receiver operator characteristic curve (AUROC), along with associated 95% CIs. We identified the top clinical features and words that appeared in clinical notes based on the LASSO and random forest models. We plotted the precision-recall curve to better understand the performance of a classification model and assessed the impact of different rule-based phenotypes.

As a way to understand the importance and potential impact of accurate phenotyping, we performed an illustrative association analysis, evaluating the relationship between vaccination status and the following hospital outcome metrics: length of stay, intensive care unit (ICU) utilization, and in-hospital mortality. These were chosen, since they are standard quality metrics for operational purposes. We regressed each outcome onto vaccination status. We used a log-linear model for length of stay and used logistic regression for ICU utilization and in-hospital mortality. Each regression was performed by using the full cohort and compared to a model that only included patients who were determined to have been hospitalized due to COVID-19. We also tested for an interaction between vaccination status and the cause of hospitalization. We emphasize that these were illustrative analyses, and they were not meant to infer any causal effects of vaccination but rather to illustrate the importance of using cause-specific phenotyping for relevant COVID-19 outcomes.

All work was performed in R version 4.1.2 (R Foundation for Statistical Computing) [20] and Python version 3.9.1 (Python Software Foundation) [21]. The processing code is available in our GitLab (GitLab Inc) [22].

Results

Patient Characteristics

In total, we reviewed the charts of 630 patients who were admitted and tested positive for SARS-CoV-2. After excluding patients younger than 18 years and patients with privacy restrictions, our data set included 586 unique patients who were hospitalized and had tested positive for SARS-CoV-2. Of these, 224 (38.2%) were determined, through clinician review, to have been hospitalized for reasons other than COVID-19. During their assessments, our chart reviewers noted that it was often readily apparent which hospitalizations were attributable to COVID-19 and which were not.

Characteristics, by admission cause, are shown in Table 1. Compared with patients hospitalized for indications other than COVID-19, patients hospitalized due to COVID-19 were, on average, older (age: mean 62.7 years vs mean 51.9 years; SMD 0.587), and their admissions were more commonly labeled as emergency admissions (346/362, 95.6% vs 165/224, 73.7%; SMD 0.641). Furthermore, patients hospitalized due to COVID-19 were substantially more likely to receive COVID-19 therapies, including steroids (233/362, 64.4% vs 54/224, 24.1%; SMD 0.887) and the antiviral agent remdesivir (247/362, 68.2% vs 55/224, 24.6%; SMD 0.974), during their hospitalization. Patients hospitalized due to COVID-19 had lower lymphocyte counts on average compared with those of patients hospitalized for reasons other than COVID-19. Normal levels of C-reactive protein and the lack of dimerized plasmin fragment D (D-dimer) testing were associated with hospitalizations for reasons other than COVID-19.

Table . Cohort description.

Characteristics	Hospitalized due to COVID-19			Standardized mean difference
	No (n=224)	Yes (n=362)	Total (N=586)	
Sex (female), n (%)	120 (53.6)	181 (50)	301 (51.4)	0.072
Age (years), mean	51.9	62.7	58.6	0.587
Patient outcome at discharge, n (%)				0.169
Dead	18 (8)	39 (10.8)	57 (9.7)	
Home	176 (78.6)	258 (71.3)	434 (74.1)	
Other facility	30 (13.4)	65 (18)	95 (16.2)	
Admission type, n (%)				0.641
Emergency admission	165 (73.7)	346 (95.6)	511 (87.2)	
Routine elective admission	24 (10.7)	4 (1.1)	28 (4.8)	
Urgent admission	35 (15.6)	12 (3.3)	47 (8)	
Transfer to intensive care unit, n (%)	45 (20.1)	78 (21.5)	123 (21)	0.036
Encounter type, n (%)				0.181
Emergency	2 (0.9)	1 (0.3)	3 (0.5)	
Emergency to inpatient	180 (80.4)	314 (86.7)	494 (84.3)	
Inpatient	31 (13.8)	35 (9.7)	66 (11.3)	
Observation stay	11 (4.9)	12 (3.3)	23 (3.9)	
Race and ethnicity, n (%)				0.168
Hispanic	21 (9.4)	20 (5.5)	41 (7)	
Non-Hispanic Black	106 (47.3)	175 (48.3)	281 (48)	
Non-Hispanic White	90 (40.2)	152 (42)	242 (41.3)	
Non-Hispanic Asian	7 (3.1)	14 (3.9)	21 (3.6)	
Other races	0 (0)	1 (0.3)	1 (0.2)	
Length of stay (days), mean	10.2	9.9	10	0.026
BMI, n (%)				0.203
Missing	9 (4)	9 (2.5)	18 (3.1)	
Normal	65 (29)	89 (24.6)	154 (26.3)	
Obese	85 (37.9)	147 (40.6)	232 (39.6)	
Overweight	60 (26.8)	98 (27.1)	158 (27)	
Underweight	5 (2.2)	19 (5.2)	24 (4.1)	
Raw payer type value, n (%)				0.305
Private	102 (45.5)	180 (49.7)	282 (48.1)	
Public	88 (39.3)	144 (39.8)	232 (39.6)	
Self-pay	21 (9.4)	9 (2.5)	30 (5.1)	
Other	13 (5.8)	29 (8)	42 (7.2)	
Vaccinated against COVID-19, n (%)	113 (50.4)	178 (49.2)	291 (49.7)	0.026
Comorbidities, n (%)				
Surgery	200 (89.3)	302 (83.4)	502 (85.7)	0.171
Cancer	29 (12.9)	45 (12.4)	74 (12.6)	0.015
Cardiovascular	75 (33.5)	146 (40.3)	221 (37.7)	0.142
Hypertension	73 (32.6)	151 (41.7)	224 (38.2)	0.19

Characteristics			Hospitalized due to COVID-19			Standardized mean difference
			No (n=224)	Yes (n=362)	Total (N=586)	
	Chronic liver disease		30 (13.4)	46 (12.7)	76 (13)	0.02
	Chronic obstructive pulmonary disease		21 (9.4)	50 (13.8)	71 (12.1)	0.139
	Asthma		18 (8)	39 (10.8)	57 (9.7)	0.094
	Chronic renal disease		44 (19.6)	111 (30.7)	155 (26.5)	0.256
	Diabetes		45 (20.1)	103 (28.5)	148 (25.3)	0.196
Medications, n (%)						
	Bronchodilator		44 (19.6)	159 (41.2)	193 (32.9)	0.481
	Steroid		54 (24.1)	233 (64.4)	287 (49)	0.887
	Anticoagulant antiplatelet		121 (54)	284 (78.5)	405 (69.1)	0.535
	Diuretic		60 (26.8)	131 (36.2)	191 (32.6)	0.203
	Cough suppressant		44 (19.6)	162 (44.8)	206 (35.2)	0.558
	Paralytic		10 (4.5)	30 (8.3)	40 (6.8)	0.157
	Expectorant		14 (6.3)	56 (15.5)	70 (11.9)	0.3
	Remdesivir		55 (24.6)	247 (68.2)	302 (51.5)	0.974
	Inhaled steroid		24 (10.7)	42 (11.6)	66 (11.3)	0.028
Laboratory tests, n (%)						
	Absolute lymphocyte count					0.345
		High	1 (0.4)	2 (0.6)	3 (0.5)	
		Low	12 (5.4)	47 (13)	59 (10.1)	
		Normal	23 (10.3)	59 (16.3)	82 (14)	
		Not taken	188 (83.9)	254 (70.2)	442 (75.4)	
	Lymphocyte count					0.528
		Low	17 (7.6)	71 (19.6)	88 (15)	
		Normal	131 (58.5)	233 (64.4)	364 (62.1)	
		Not taken	76 (33.9)	56 (15.5)	132 (22.5)	
		High	0 (0)	2 (0.6)	2 (0.3)	
	C-reactive protein					0.602
		High	62 (27.7)	203 (56.1)	265 (45.2)	
		Normal	11 (4.9)	9 (2.5)	20 (3.4)	
		Not taken	151 (67.4)	150 (41.4)	301 (51.4)	
	Ferritin					0.361
		High	39 (17.4)	107 (29.6)	146 (24.9)	
		Low	2 (0.9)	3 (0.8)	5 (0.9)	
		Normal	17 (7.6)	44 (12.2)	61 (10.4)	
		Not taken	166 (74.1)	208 (57.5)	374 (63.8)	
	D-dimer[a]					1.187
		High	19 (8.5)	117 (32.3)	136 (23.2)	
		Normal	36 (16.1)	156 (43.1)	192 (32.8)	
		Not taken	169 (75.4)	89 (24.6)	258 (44)	
	Procalcitonin					0.524
		High	4 (1.8)	22 (6.1)	26 (4.4)	

Characteristics		Hospitalized due to COVID-19			Standardized mean difference
		No (n=224)	Yes (n=362)	Total (N=586)	
	Missing	208 (92.9)	268 (74)	476 (81.2)	
	Normal	12 (5.4)	72 (19.9)	84 (14.3)	

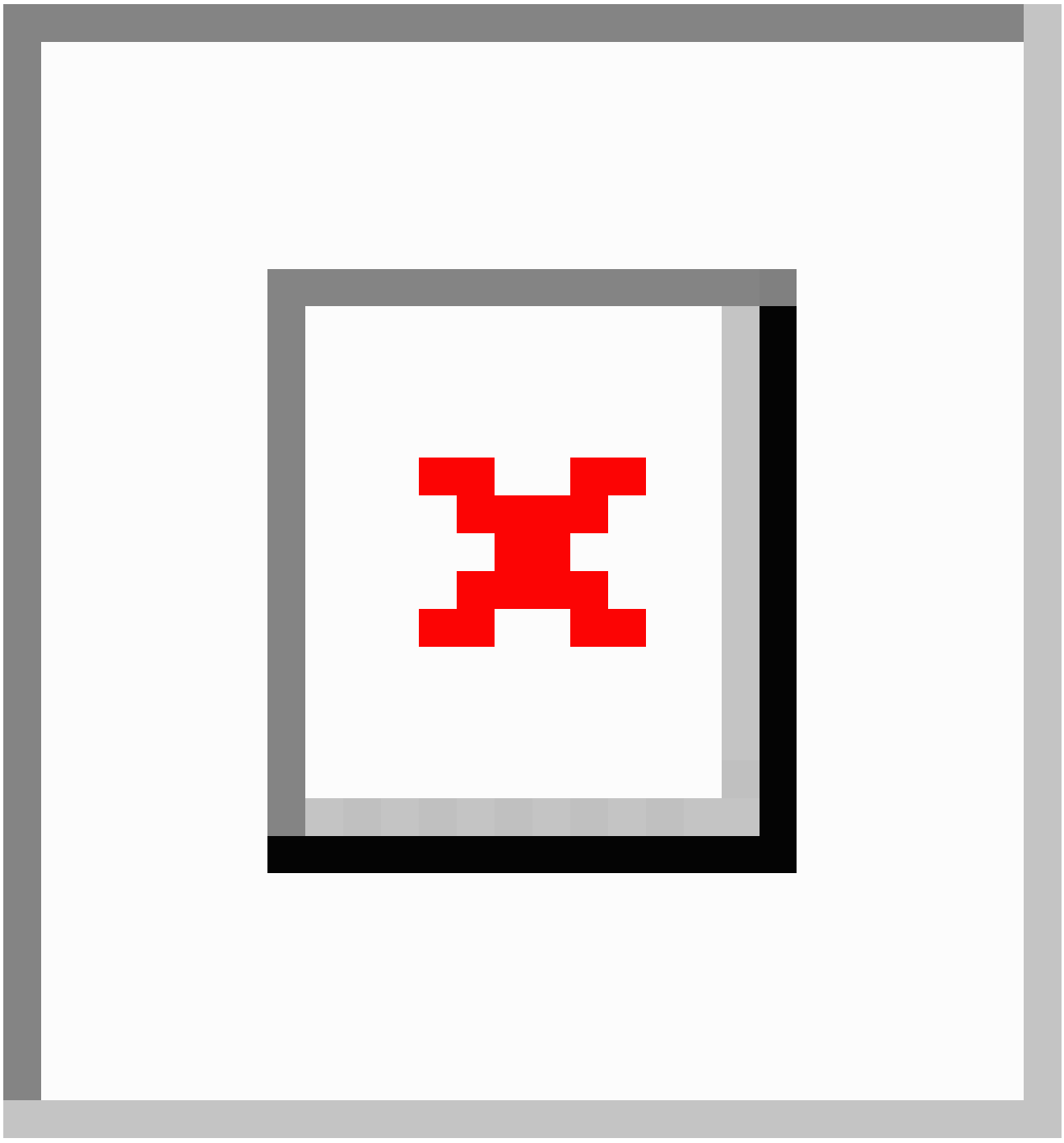
[a]D-dimer: dimerized plasmin fragment D.

Performance of Classification Models

After tokenizing words and removing terms with fewer than 50 occurrences, our models included 7953 unique terms. There was minimal difference between the LASSO and random forest models. The random forest model based solely on clinical notes, the one based solely on structured data elements, and the one that used both clinical notes and structured data elements had AUROCs of 0.882 (95% CI 0.85-0.909), 0.829 (95% CI 0.794-0.864), and 0.890 (95% CI 0.864-0.916), respectively. The LASSO model based solely on clinical notes (AUROC=0.894, 95% CI 0.868-0.920) had better discrimination than the LASSO model based solely on structured data elements (AUROC=0.841, 95% CI 0.809-0.874; *P*<.001). The LASSO model using both clinical notes and structured data elements (AUROC=0.893, 95% CI 0.868-0.919) had similar discrimination to that of the LASSO model based solely on clinical notes (*P*=.91).

Next, we examined the top structured data elements and terms in each model (Figure 1). Highly predictive data elements and words corresponded to patient characteristics with large SMDs (Table 1). Words that are reflective of hospitalization due to COVID-19 have positive coefficients, while words reflective of hospitalization for other reasons have negative coefficients. Terms reflective of COVID-19–specific hospitalization were related to the care of patients with COVID-19, such as “remdesivir” and “dexamethason.” Other structured elements related to the likelihood of being hospitalized for COVID-19 included receipt of steroids, low lymphocyte counts, and underweight BMIs. Terms reflective of hospitalizations due to indications other than COVID-19 included strings that may be related to surgical procedures (eg, “surgic” for “surgical” or “dress” for “dressing”). For structured data elements, a lack of D-dimer collection and low ferritin levels were most commonly associated with admissions for reasons other than COVID-19. Similar features were identified from the random forest model (Figure S1 in Multimedia Appendix 1).

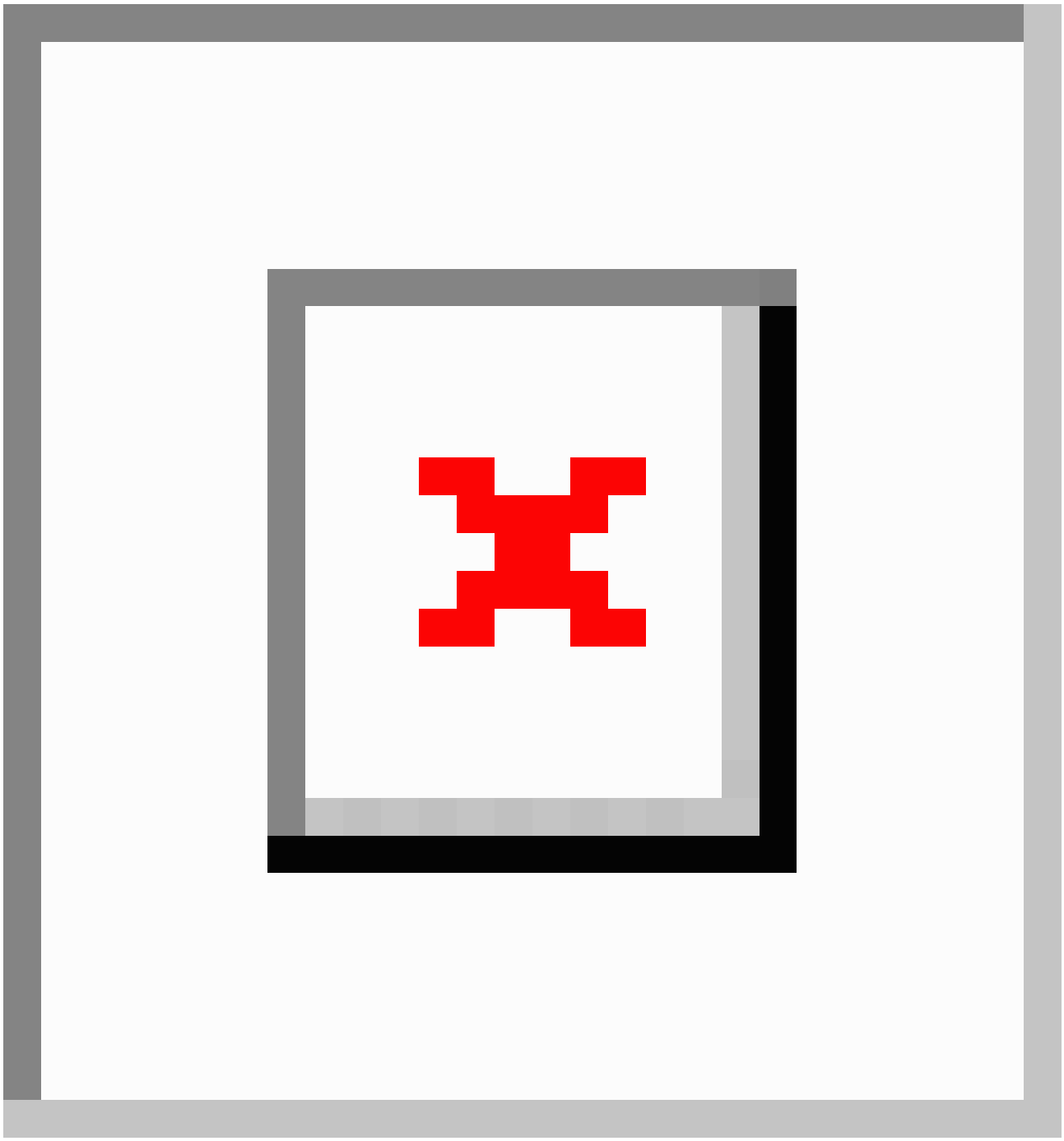
Figure 1. The top regression coefficients from the LASSO models, as reflective of variable importance for (A) the model using just structured data elements and (B) the model using just clinical notes. Values greater than 0 indicate that the feature has a positive association with hospitalization due to COVID-19, while values less than 0 indicate that a feature has a negative association.



Impact of Correct Classification

In order to assess the performance of a computable phenotype–based decision rule, we examined the precision-recall curve of the different models (Figure 2). For example, a rule that maintains a sensitivity of 90% (ie, one that would capture 90% of all patients hospitalized due to COVID-19) resulted in positive predictive values of 76%, 82%, and 84% and corresponding F_1-scores of 0.824, 0.858, and 0.869 based on structured data elements, clinical notes, and their combination, respectively. To illustrate the impact of these differences, we considered the impact of implementing each of these phenotypes at a 90% sensitivity to classify patients during the January Omicron wave. Within our health system, 1378 people were hospitalized and tested positive for SARS-CoV-2. Based on our analyses, using the LASSO-based phenotype that incorporates structured data, clinical notes, or their combination would result in approximately 244, 165, and 142 false positives, respectively.

Figure 2. Precision-recall (positive predictive value and sensitivity) curve for the different classification algorithms. This illustrates the trade-off between the identification of patients hospitalized due to COVID-19 (x-axis: sensitivity) and the accuracy of that capture (y-axis: positive predictive value). There is minimal difference between using just notes or notes with structured data elements. The model with only structured data elements performs notably worse in terms of positive predictive value at the same sensitivity thresholds. AUPRC: area under the precision-recall curve.



We next sought to evaluate the potential impact of different phenotyping methods on hospital outcome metrics, comparing a method that incorporates the reason for hospitalization versus one that does not. We used a regression analysis to assess the marginal relationship. As a use case, we evaluated associations between vaccine status and the following three hospital outcome metrics: length of stay, risk of ICU utilization, and in-hospital mortality. These evaluations were performed with the following three cohorts: all hospitalized patients, those who were determined to have been hospitalized due to COVID-19, and those who tested positive for SARS-CoV-2 but were hospitalized for unrelated reasons (Table 2). For length of stay, the magnitude of the effect of vaccine status changed based on the cohort used. In the cohort of all hospitalized patients, vaccinated patients had a shorter length of stay (relative rate 0.81, 95% CI 0.71-0.93). However, when limiting the analytic cohort to patients hospitalized due to COVID-19, there was no significant difference in length of stay for vaccinated patients versus unvaccinated patients (relative rate 0.98, 95% CI 0.83-1.16; *P* value for interaction<.001). We found similar patterns in analyses of other in-hospital outcomes; vaccination was associated with reduced risks of ICU utilization and in-hospital mortality among patients hospitalized for reasons other than COVID-19 when compared to those among patients hospitalized

due to COVID-19. Effects were robust to adjustment for age (Table S2 in Multimedia Appendix 1). These results illustrate the impact of selecting the correct cohort for analysis and the potential ramifications of using a cohort in which the reason for hospitalization has not been determined.

Table . Marginal association between vaccine status[a] and outcome metrics, unadjusted for age.

Outcome	Full cohort	Hospitalized due to COVID-19	Hospitalization unrelated to COVID-19	*P* value[b]
Length of stay, relative rate (95% CI)	0.81 (0.71-0.93)	0.98 (0.83-1.16)	0.59 (0.47-0.74)	<.001
ICU[c] utilization, odds ratio (95% CI)	1.04 (0.70-1.56)	1.25 (0.75-2.07)	0.77 (0.40-1.49)	.26
Mortality, odds ratio (95% CI)	1.02 (0.59-1.78)	1.45 (0.74-2.88)	0.48 (0.16-1.29)	.08

[a]Unvaccinated patients are the reference group.

[b]*P* value is for hospitalization due to COVID-19 versus hospitalization unrelated to COVID-19.

[c]ICU: intensive care unit.

Discussion

Principal Findings

Due to the public health importance of the accurate identification of COVID-19–related hospitalizations, there is a need for methods and computable phenotypes to identify hospital admissions in which the primary cause is COVID-19 [23]. We used machine learning methods and a physician chart review to develop a classification algorithm for hospitalization due to COVID-19. We found that 38.2% (224/586) of patients who were hospitalized at our institution during the Omicron wave and tested positive for SARS-CoV-2 infection were hospitalized for reasons other than COVID-19. These findings are in line with other recent studies, which found that an average of 26% of hospitalized patients with a positive SARS-CoV-2 test result had a primary indication for hospitalization that was unrelated to COVID-19 [14]. We found that a model based on clinical notes performed better than one based solely on structured EHR data elements. This work has important implications for retrospective analyses using EHR data to assess outcomes related to COVID-19, including vaccine effectiveness and health system capacity [24].

Prior work by Lynch and colleagues [25] evaluated the utility of *ICD-10* codes for COVID-19 diagnosis in inpatient, outpatient, emergency care, and urgent care settings during time periods across the pandemic; using a weighted, random sample of 1500 records from the Department of Veterans Affairs, they found that the COVID-19 *ICD-10* code (U07.1) had a relatively low positive predictive value across settings and time periods. These findings highlight the need for additional contextual data to identify acute cases of COVID-19. The Consortium for Clinical Characterization of COVID-19 by EHR (4CE) conducted a similar study of EHR data from 12 clinical sites to identify combinations of structured data elements to generate a reliable computable phenotype for hospitalization due to COVID-19, with a reported AUROC of 0.903 [26]. Similarly, we derived an AUROC of 0.841 based solely on structured data elements; however, we also found that that inclusion of clinical notes significantly improved the performance of the classification model (AUROC=0.893; *P*<.001). This result is not surprising, as the clinical narrative often includes important nuance, and as our chart reviewers noted, it was often readily apparent which hospitalizations were attributable to COVID-19 and which were not. Of note, chart reviewers in our study classified hospitalizations that were indirectly due to SARS-CoV-2 infection, such as those due to COVID-19–related weakness or delirium, as hospitalizations due to COVID-19, which could partly explain the observed difference in discriminatory ability between our study and the study conducted by the 4CE.

By using the TF-IDF approach in conjunction with LASSO regression, we identified both individual terms and the direction of the association between each term and the hospitalization indication. Although the TF-IDF approach is a simple natural language processing (NLP) approach, it is also very scalable, interpretable, and implementable. Our results highlight the power of even simple natural language models. The terms that best predicted hospitalizations due to COVID-19 included common descriptors that were used in the clinical care of patients with COVID-19, such as “hypox” (likely shortened from “hypoxia” or “hypoxic”), or COVID-19 therapies like remdesivir. Conversely, the terms that were not associated with hospitalizations due to COVID-19 included words related to surgery—a common indication for hospital admission that is generally unrelated to SARS-CoV-2 infection.

To help contextualize our results, we also assessed the real-world impact of using an accurate phenotype for COVID-19–specific hospitalization. In studying hospitalized patients with COVID-19, the simplest analysis would be to include all patients with a COVID-19–positive test result. As our illustrative analysis showed, when using this full but heterogeneous cohort, the results suggested that vaccination status is associated with a shorter length of stay. However, when we limited the analysis to only include patients who were identified as having been hospitalized due to COVID-19 (ie, people with symptoms of COVID-19), the analysis indicated that vaccines are not associated with a shorter length of stay. We interpreted these data as indicating that, conditional on someone being sick enough to be hospitalized due to COVID-19, vaccines provide no additional benefit in terms of the length of

hospitalization. Similar patterns were found for other hospital outcome metrics. Although this analysis was not intended to be a causal analysis, it did illustrate how the use of accurately classified cohorts is important for the calculation of standard outcome metrics and likely impacts other related association analyses.

More broadly, this work highlights the importance and challenge of phenotyping cause-specific events. Although there is rich literature on computable phenotypes, most of this literature is geared toward the identification of chronic diseases (eg, presence of asthma). However, few computable phenotypes have focused on cause-specific events (eg, asthma exacerbation). Such cause-specific phenotypes often exhibit poor specificity and can require algorithms that are more complex than those required for chronic conditions. As this work shows, and as suggested by others, NLP-based phenotyping approaches are becoming more common, and further comparisons between NLP approaches and other methods will be needed to determine whether using text data can improve cause-specific phenotypes.

Although our study used rigorous methods, there are some key limitations. First and most notably, our findings are primarily illustrative and may not represent a generalizable algorithm for phenotyping COVID-19–specific hospitalizations. This study was conducted across a single hospital system, and it may not be reflective of practices at other institutions. Importantly, we would not expect our specific phenotype algorithm to be generalizable to other institutions. Second, we only looked at 1 period of time, namely the January 2022 Omicron wave; however, there are documented differences in the rate of hospitalization and positive test results over the course of the pandemic, and our models may not accurately reflect distinguishing factors in other waves. Third, another limitation is that, given the time constraints of chart reviews, we were only able to analyze a relatively small sample. In particular, the small sample size limited our ability to apply more sophisticated NLP-based approaches, such as the use of n-grams.

Conclusions

Overall, our results show that a sizable number of people who were hospitalized and tested positive for SARS-CoV-2 were hospitalized for reasons other than COVID-19. The conflation of these individuals can impact our understanding of hospital outcome metrics. We constructed a strong classification model that can be used as a computable phenotype to distinguish patients who were hospitalized due to COVID-19 from those who incidentally tested positive for SARS-CoV-2 but were hospitalized for other reasons. Moreover, we found that while structured data elements are useful in constructing such a phenotype, clinical notes had a higher positive predictive value than that of structured data elements alone. Future work should seek to explore the generalizability of such phenotypes across institutions and different waves of the COVID-19 pandemic.

Acknowledgments

This work was supported by Food and Drug Administration Broad Agency Announcement (FDA BAA) 75F40121C00158 (principal investigator: BAG). We thank Mike Chrestensen for support in extracting the clinical notes for this study.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplemental materials regarding variable descriptions, top data elements, and the association between vaccine status and outcome metrics.
[DOCX File, 244 KB - medinform_v11i1e46267_app1.docx]

References

1. Mehrotra DV, Janes HE, Fleming TR, Annunziato PW, Neuzil KM, Carpp LN, et al. Clinical endpoints for evaluating efficacy in COVID-19 vaccine trials. Ann Intern Med 2021 Feb;174(2):221-228. [doi: 10.7326/M20-6169] [Medline: 33090877]
2. Centers for Disease Control and Prevention. Science brief: indicators for monitoring COVID-19 community levels and making public health recommendations. URL: www.cdc.gov/coronavirus/2019-ncov/science/science-briefs/indicators-monitoring-community-levels.html [accessed 2023-01-19]
3. Fillmore NR, La J, Zheng C, Doron S, Do NV, Monach PA, et al. The COVID-19 hospitalization metric in the pre- and postvaccination eras as a measure of pandemic severity: a retrospective, nationwide cohort study. Infect Control Hosp Epidemiol 2022 Dec;43(12):1767-1772. [doi: 10.1017/ice.2022.13] [Medline: 35012694]
4. León TM, Dorabawila V, Nelson L, Lutterloh E, Bauer UE, Backenson B, et al. COVID-19 cases and hospitalizations by COVID-19 vaccination status and previous COVID-19 diagnosis - California and New York, May-November 2021. MMWR Morb Mortal Wkly Rep 2022 Jan 28;71(4):125-131. [doi: 10.15585/mmwr.mm7104e1] [Medline: 35085222]
5. Fall A, Eldesouki RE, Sachithanandham J, Morris CP, Norton JM, Gaston DC, et al. A quick displacement of the SARS-CoV-2 variant Delta with Omicron: unprecedented spike in COVID-19 cases associated with fewer admissions and

comparable upper respiratory viral loads. medRxiv. Preprint posted online on January 28, 2022. . [doi: 10.1101/2022.01.26.22269927] [Medline: 35118480]
6. Stowe J, Andrews N, Kirsebom F, Ramsay M, Bernal JL. Effectiveness of COVID-19 vaccines against Omicron and Delta hospitalisation, a test negative case-control study. Nat Commun 2022 Sep 30;13(1):5736. [doi: 10.1038/s41467-022-33378-7] [Medline: 36180428]
7. Havers FP, Patel K, Whitaker M, Milucky J, Reingold A, Armistead I, et al. Laboratory-confirmed COVID-19-associated hospitalizations among adults during SARS-CoV-2 Omicron BA.2 variant predominance - COVID-19-Associated Hospitalization Surveillance Network, 14 States, June 20, 2021-May 31, 2022. MMWR Morb Mortal Wkly Rep 2022 Aug 26;71(34):1085-1091. [doi: 10.15585/mmwr.mm7134a3] [Medline: 36006841]
8. Taylor CA, Whitaker M, Anglin O, Milucky J, Patel K, Pham H, et al. COVID-19-associated hospitalizations among adults during SARS-CoV-2 Delta and Omicron variant predominance, by race/ethnicity and vaccination status - COVID-NET, 14 States, July 2021-January 2022. MMWR Morb Mortal Wkly Rep 2022 Mar 25;71(12):466-473. [doi: 10.15585/mmwr.mm7112e2] [Medline: 35324880]
9. Hilal W, Chislett MG, Snider B, McBean EA, Yawney J, Gadsden SA. Use of AI to assess COVID-19 variant impacts on hospitalization, ICU, and death. Front Artif Intell 2022 Nov 30;5:927203. [doi: 10.3389/frai.2022.927203] [Medline: 36530359]
10. de Jesús Ascencio-Montiel I, Ovalle-Luna OD, Rascón-Pacheco RA, Borja-Aburto VH, Chowell G. Comparative epidemiology of five waves of COVID-19 in Mexico, March 2020-August 2022. BMC Infect Dis 2022 Oct 31;22(1):813. [doi: 10.1186/s12879-022-07800-w] [Medline: 36316634]
11. Chu VT, Schwartz NG, Donnelly MAP, Chuey MR, Soto R, Yousaf AR, et al. Comparison of home antigen testing with RT-PCR and viral culture during the course of SARS-CoV-2 infection. JAMA Intern Med 2022 Jul 1;182(7):701-709. [doi: 10.1001/jamainternmed.2022.1827] [Medline: 35486394]
12. Chin ET, Huynh BQ, Chapman LAC, Murrill M, Basu S, Lo NC. Frequency of routine testing for COVID-19 in high-risk healthcare environments to reduce outbreaks. medRxiv. Preprint posted online on September 9, 2020. . [doi: 10.1101/2020.04.30.20087015] [Medline: 32511523]
13. Rickman HM, Rampling T, Shaw K, Martinez-Garcia G, Hail L, Coen P, et al. Nosocomial transmission of coronavirus disease 2019: a retrospective study of 66 hospital-acquired cases in a London teaching hospital. Clin Infect Dis 2021 Feb 16;72(4):690-693. [doi: 10.1093/cid/ciaa816] [Medline: 32562422]
14. Klann JG, Strasser ZH, Hutch MR, Kennedy CJ, Marwaha JS, Morris M, et al. Distinguishing admissions specifically for COVID-19 from incidental SARS-CoV-2 admissions: national retrospective electronic health record study. J Med Internet Res 2022 May 18;24(5):e37931. [doi: 10.2196/37931] [Medline: 35476727]
15. Hurst JH, Liu Y, Maxson PJ, Permar SR, Boulware LE, Goldstein BA. Development of an electronic health records datamart to support clinical and population health research. J Clin Transl Sci 2020 Jun 23;5(1):e13. [doi: 10.1017/cts.2020.499] [Medline: 33948239]
16. Uther W, Mladenić D, Ciaramita M, Berendt B, Kołcz A, Grobelnik M, et al. TF–IDF. In: Sammut C, Webb GI, editors. Encyclopedia of Machine Learning. Boston, MA: Springer; 2011. [doi: 10.1007/978-0-387-30164-8]
17. NLTK. Natural language toolkit. URL: www.nltk.org/ [accessed 2022-10-30]
18. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Series B Stat Methodol 1996 Jan;58(1):267-288. [doi: 10.1111/j.2517-6161.1996.tb02080.x]
19. Breiman L. Random forests. Mach Learn 2001 Oct;45(1):5-32. [doi: 10.1023/A:1010933404324]
20. R Foundation for Statistical Computing. The R project for statistical computing. 2021. URL: www.R-project.org/ [accessed 2023-07-28]
21. Python Software Foundation. Python release Python 3.9.1. URL: www.python.org/downloads/release/python-391/ [accessed 2023-01-26]
22. GitLab. Using nlp to identify computable phenotype of patients hospitalized because of COVID-19. URL: gitlab.com/changalice980/using-nlp-to-identify-computable-phenotype-of-patients-hospitalized-because-of-covid-19 [accessed 2023-07-28]
23. U.S. Department of Health and Human Services. COVID-19 guidance for hospital reporting and FAQs for hospitals, hospital laboratory, and acute care facility data reporting. URL: www.hhs.gov/sites/default/files/covid-19-faqs-hospitals-hospital-laboratory-acute-care-facility-data-reporting.pdf [accessed 2023-07-17]
24. Satterfield BA, Dikilitas O, Kullo IJ. Leveraging the electronic health record to address the COVID-19 pandemic. Mayo Clin Proc 2021 Jun;96(6):1592-1608. [doi: 10.1016/j.mayocp.2021.04.008] [Medline: 34088418]
25. Lynch KE, Viernes B, Gatsby E, DuVall SL, Jones BE, Box TL, et al. Positive predictive value of COVID-19 ICD-10 diagnosis codes across calendar time and clinical setting. Clin Epidemiol 2021 Oct 27;13:1011-1018. [doi: 10.2147/CLEP.S335621] [Medline: 34737645]
26. Hong C, Zhang HG, L'Yi S, Weber G, Avillach P, Tan BWQ, et al. Changes in laboratory value improvement and mortality rates over the course of the pandemic: an international retrospective cohort study of hospitalised patients infected with SARS-CoV-2. BMJ Open 2022 Jun 23;12(6):e057725. [doi: 10.1136/bmjopen-2021-057725] [Medline: 35738646]

Abbreviations

4CE: Consortium for Clinical Characterization of COVID-19 by EHR
AUROC: area under the receiver operator characteristic curve
D-dimer: dimerized plasmin fragment D
DUHS: Duke University Health System
EHR: electronic health record
***ICD-10*:** *International Classification of Diseases, 10th Revision*
ICU: intensive care unit
LASSO: least absolute shrinkage and selection operator
NLP: natural language processing
SMD: standardized mean difference
TF-IDF: term frequency–inverse document frequency

Edited by A Benis; submitted 07.02.23; peer-reviewed by ME Chatzimina, Q Chen, W Ip; revised version received 19.05.23; accepted 17.06.23; published 22.08.23.

Please cite as:
Chang F, Krishnan J, Hurst JH, Yarrington ME, Anderson DJ, O'Brien EC, Goldstein BA
Comparing Natural Language Processing and Structured Medical Data to Develop a Computable Phenotype for Patients Hospitalized Due to COVID-19: Retrospective Analysis
JMIR Med Inform 2023;11:e46267
URL: https://medinform.jmir.org/2023/1/e46267
doi:10.2196/46267

Original Paper

Unique Device Identification–Based Linkage of Hierarchically Accessible Data Domains in Prospective Surgical Hospital Data Ecosystems: User-Centered Design Approach

Karol Kozak[1], Prof Dr; André Seidel[2], PhD; Nataliia Matvieieva[2], MA; Constanze Neupetsch[2,3], MA; Uwe Teicher[2], PhD; Gordon Lemme[2], MSc; Anas Ben Achour[2], MSc; Martin Barth[4], MA; Steffen Ihlenfeldt[2,5], Prof Dr; Welf-Guntram Drossel[2,3], Prof Dr

[1]Center for Evidence-Based Healthcare, Technische Universität Dresden, Dresden, Germany

[2]Fraunhofer Institute for Machine Tools and Forming Technology IWU, Dresden, Germany

[3]Professorship for Adaptronics and Lightweight Design in Production, Technische Universität Chemnitz, Chemnitz, Germany

[4]Fraunhofer Institute for Ceramic Technologies and Systems IKTS, Dresden, Germany

[5]Chair of Machine Tools Development and Adaptive Controls, Technische Universität Dresden, Dresden, Germany

Corresponding Author:
Uwe Teicher, PhD
Fraunhofer Institute for Machine Tools and Forming Technology IWU
Pforzheimer Straße 7a
Dresden, 01187
Germany
Phone: 49 351 4772 ext 2620
Email: uwe.teicher@iwu.fraunhofer.de

Abstract

Background: The electronic health record (EHR) targets systematized collection of patient-specific, electronically stored health data. The EHR is an evolving concept driven by ongoing developments and open or unclear legal issues concerning medical technologies, cross-domain data integration, and unclear access roles. Consequently, an interdisciplinary discourse based on representative pilot scenarios is required to connect previously unconnected domains.

Objective: We address cross-domain data integration including access control using the specific example of a unique device identification (UDI)–expanded hip implant. In fact, the integration of technical focus data into the hospital information system (HIS) is considered based on surgically relevant information. Moreover, the acquisition of social focus data based on mobile health (mHealth) is addressed, covering data integration and networking with therapeutic intervention and acute diagnostics data.

Methods: In addition to the additive manufacturing of a hip implant with the integration of a UDI, we built a database that combines database technology and a wrapper layer known from extract, transform, load systems and brings it into a SQL database, WEB application programming interface (API) layer (back end), interface layer (rest API), and front end. It also provides semantic integration through connection mechanisms between data elements.

Results: A hip implant is approached by design, production, and verification while linking operation-relevant specifics like implant-bone fit by merging patient-specific image material (computed tomography, magnetic resonance imaging, or a biomodel) and the digital implant twin for well-founded selection pairing. This decision-facilitating linkage, which improves surgical planning, relates to patient-specific postoperative influencing factors during the healing phase. A unique product identification approach is presented, allowing a postoperative read-out with state-of-the-art hospital technology while enabling future access scenarios for patient and implant data. The latter was considered from the manufacturing perspective using the process manufacturing chain for a (patient-specific) implant to identify quality-relevant data for later access. In addition, sensor concepts were identified to use to monitor the patient-implant interaction during the healing phase using wearables, for example. A data aggregation and integration concept for heterogeneous data sources from the considered focus domains is also presented. Finally, a hierarchical data access concept is shown, protecting sensitive patient data from misuse using existing scenarios.

Conclusions: Personalized medicine requires cross-domain linkage of data, which, in turn, require an appropriate data infrastructure and adequate hierarchical data access solutions in a shared and federated data space. The hip implant is used as an example for the usefulness of cross-domain data linkage since it bundles social, medical, and technical aspects of the implantation.

It is necessary to open existing databases using interfaces for secure integration of data from end devices and to assure availability through suitable access models while guaranteeing long-term, independent data persistence. A suitable strategy requires the combination of technical solutions from the areas of identity and trust, federated data storage, cryptographic procedures, and software engineering as well as organizational changes.

(JMIR Med Inform 2023;11:e41614) doi:10.2196/41614

KEYWORDS

electronic health record; unique device identification; cyber-physical production systems; mHealth; data integration ecosystem; hierarchical data access; shell embedded role model

Introduction

Unique device identification (UDI) is a system used to identify devices within the health care supply chain based on a consistent, standardized, and unambiguous machine-readable identifier to keep track of the postmarketing performance of medical devices [1]. The performance of a hip implant, for example, cannot be evaluated without considering individual factors of the recipient [2] and the conditions of the therapeutic intervention [3]. Consequently, patient, medicine, and product have to be linked and monitored to enable a well-founded evaluation. Regardless of the still-missing legal framework conditions [4] and the existing ethical and political questions [5], digitalization of the health system is advancing [6], which is beneficial for a more holistic assessment. Either way, part of this development is the cross-domain linkage of data [7] that at least can be resolved on a patient-by-patient basis to prepare for intelligent data analysis. This requires going beyond analysis objectives to an appropriate data infrastructure, which enables different data domains to be linked while providing adequate hierarchical data access concepts. Consequently, this paper approached a framework for cross-domain cooperation and intelligent data analysis in a specific application scenario embedded in prospective digital hospital ecosystems. Linking unconnected domains inside of safe frameworks with clarified data sovereignty to enable a holistic approach to personalized treatment benefits health care, lowers risks of mistreatments, and functions as a catalyst for the optimization of medical products. This concept, based on a pilot scenario, can function as a basis to clarify unclear legal issues.

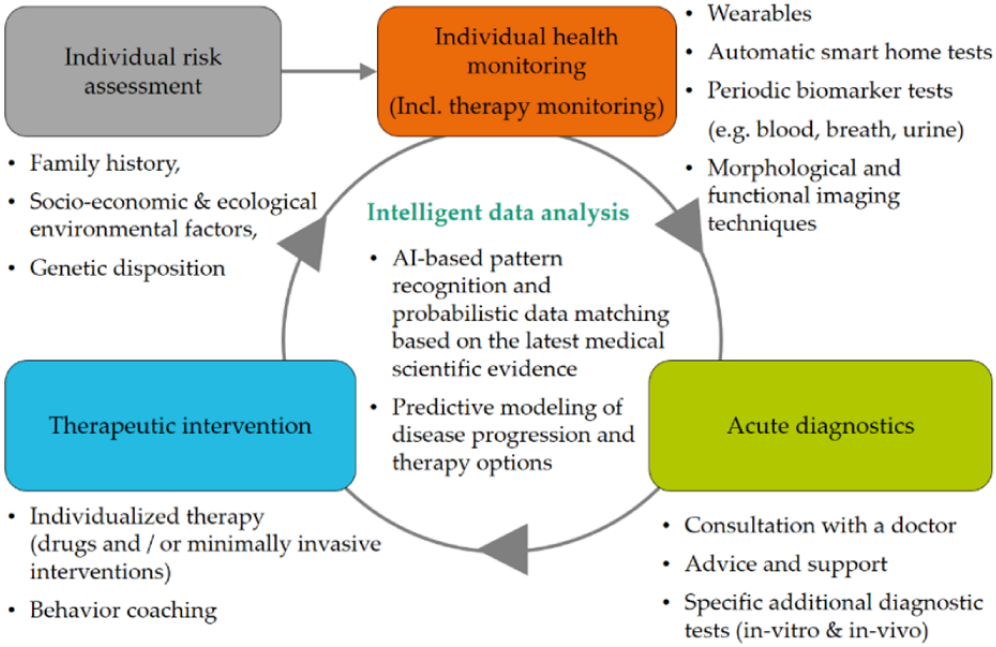
Methods

Framework Conditions for Digitization in Individualized Medicine

Intelligent Data Analysis Framework

An EHR is the systematized collection of electronically stored patient and population health data in a digital format. These records can be shared across different health care settings [8]. Records are provided through network-connected, enterprise-wide information systems or other information networks and exchanges. EHRs can include a range of data such as individual risk assessments, health monitoring, acute diagnostics, and therapeutic interventions while enabling comprehensive, intelligent data analysis, which, according to Hahn et al [9], is considered the information cycle of personalized medicine (Figure 1).

Figure 1. Information cycle of individualized medicine [9]. AI: artificial intelligence.

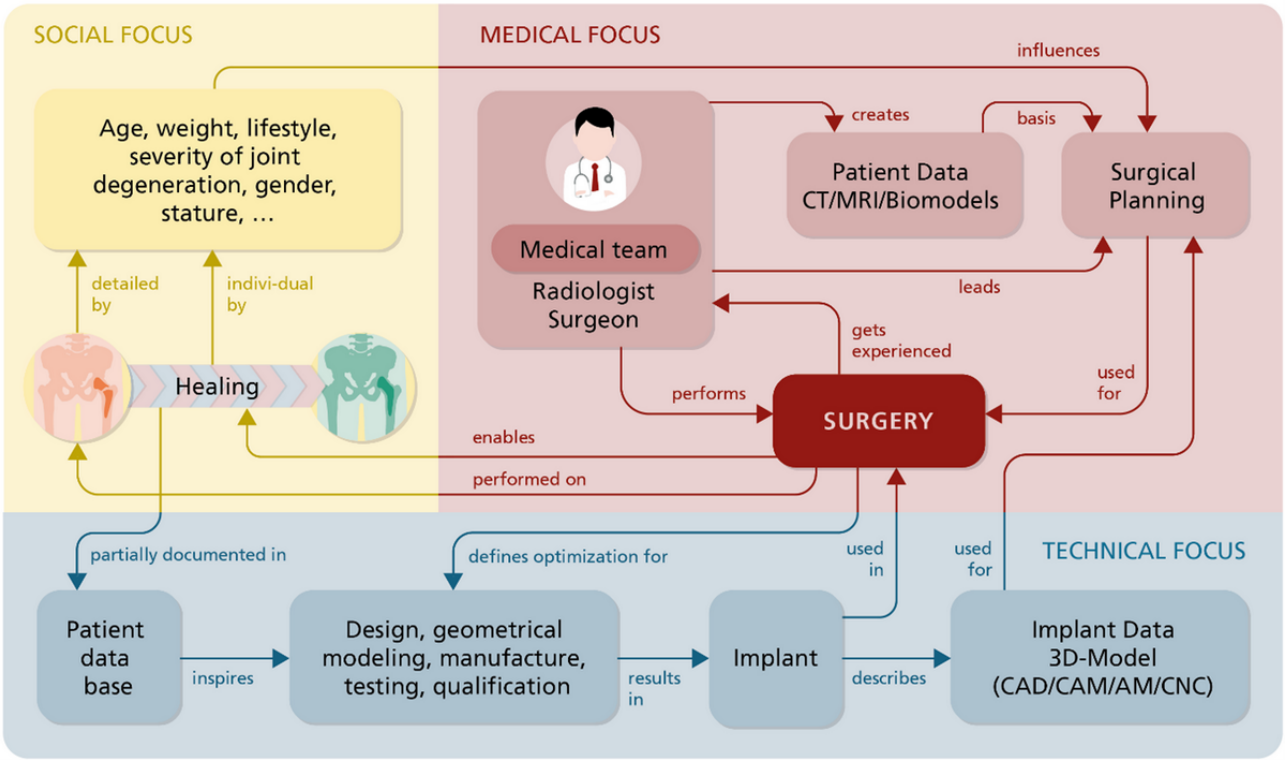


These data are stored as a patient digital twin in a centralized hospital information system (HIS) in the form of text, images (DICOM and other types), and scans. EHRs enable patients and hospitals to manage their health information in public (eg, hospital) and private environments as a personal health record (PHR). The information contained in EHRs is highly sensitive. Unintended exposure of these data threatens an intimate part of a patient's private sphere and may lead to undesirable consequences. The EHR is a communication tool that supports clinical decision-making, coordination of services (illness type, care type), evaluation of the quality and efficacy of care, research, legal protection, education, and accreditation, and regulatory processes. It is the business record of the health care system, documented in the normal course of its activities. Patients routinely review EHRs and keep PHR in their own digital archive or in patient portals (eg, at health insurance companies such as "TK-Safe," "Vivy," or "AOK-Gesundheitsnetzwerk" in Germany), given the patient is the owner of the EHR and PHR. The physician, practice, or organization is the owner of the physical medical record because it is its business record and property, and the patient owns the information in the medical record. Although the record belongs to the care facility or doctor, it is truly also the patient's information. EHRs should be released to other stakeholders only with the patient's permission or as allowed by law or via studies: public registry and ministries. PHRs are already on the market. The purpose of a PHR is to maintain good health and target outcomes; this could include daily vital signs (eg, blood pressure, heart rate), number of walking steps, amount of exercise, and calorie intake. In addition to these, information for medical use might be considered, such as blood type, allergies, pre-existing diseases, medicines the user is taking, emergency contact information, and information about the user's medical institution.

Synergy Potential in Information Linkage Using the Example of a Hip Implant

Neugebauer [10] stated that the digital transformation is promoted by the interaction of technologies that were previously perceived as independent of each other. Hahn et al [9] noted, in this context, that networking of the individual sectors and the structured use of integrated information are still pending. Moreover, Hahn et al [9] concluded that sustainable success can only be expected if the interlinking of technological and biomedical research on the one hand and clinical implementation and product development on the other hand are permanently guaranteed. Either way, this paper approached data-driven interdisciplinary research from an application-oriented perspective in an incident-based scenario (Figure 2). This example illustrates selected dependencies between social parameters, medical factors and technical aspects important for surgery, and healing, which are currently not linked sufficiently. In fact, it is a common practice to laboriously obtain this information on a case-by-case basis, which requires appropriate lead time before the operation; this is a major disadvantage in the case of emergency medical treatment. Consequently, this paper addressed this shortcoming and developed a possible solution scenario showing how this information can be linked at the EHR level. Moreover, we've shown how this information can be made accessible based on implant-inherent features while introducing a role model for access regulation and data protection.

Figure 2. Incident-based networking of social aspects, medical factors, and technical aspects. AM: additive manufacturing; CAD: computer-aided design; CAM: computer-aided manufacturing; CNC: computer numerical control; CT: computed tomography; MRI: magnetic resonance imaging.

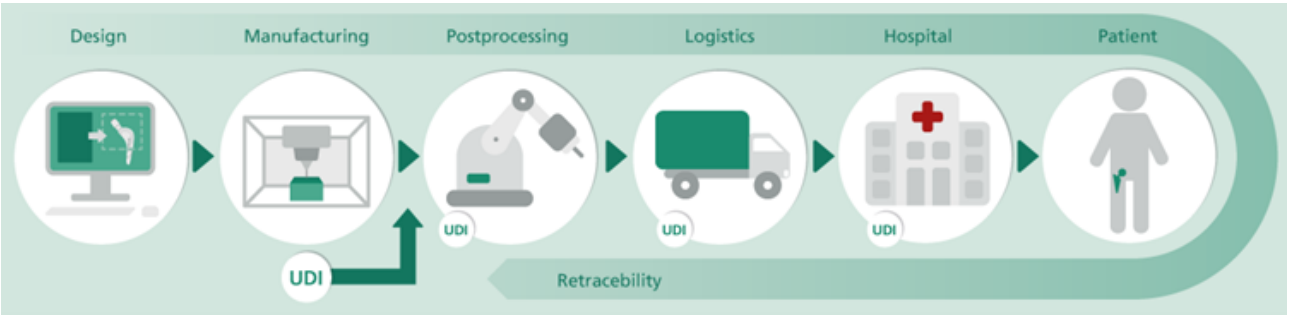


Data-Driven Networking of Information

Regulatory Demands

In the medical technology industry, quality assurance is a particular focus due to the stringent regulations. With a changeover period, the European Medical Device Regulation (MDR) EU 2017/745 came into effect in May 2021 and replaced the European directives on medical devices. The regulation obligates manufacturers to mark medical devices that are marketed in the European Union with unique codes. The main objective for the introduction of these codes is to increase patient safety. Unique product identification prevents confusion of medical devices and makes counterfeiting more difficult. The markings are implemented by the UDI system. The UDI enables tracking of medical devices, for example, from the last step of postprocessing in manufacturing where the marking is applied to the component (eg, by laser engraving). Nevertheless, the medical products can only be tracked from this point through the logistics process to the hospital. Figure 3 illustrates selected phases of a medical device life cycle covering design, manufacturing, postprocessing, logistics, and the union with the patient. Unfortunately, doubtless identification of an implant after implantation is impossible with the UDI system, which largely excludes proof of quality and originality after implantation. Across industries, product piracy is a major problem, and the estimated economic damage has been increasing over the years [11]. In addition to the generally valid comments on product piracy and its consequences, other aspects have to be considered in the field of medical technology. After the publication of the "Implant Files" in 2018 by a group of investigative journalists, the problem of defective medical devices became a sociopolitical issue [12].

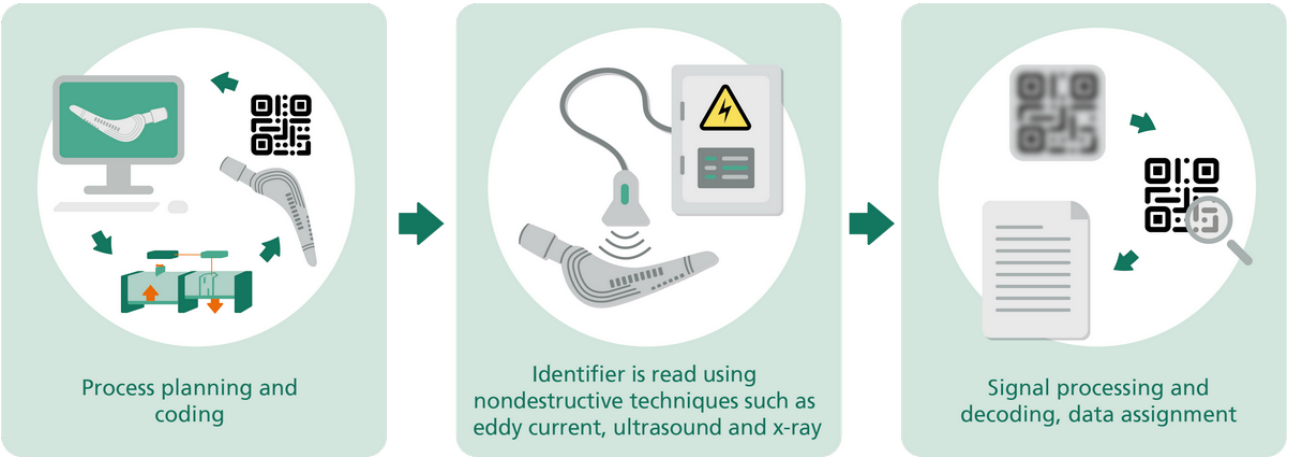
Figure 3. Retraceability of a medical device with a unique device identification (UDI).



Component-Inherent Identifier–Based Data Access

Inherent markings of implants could be a solution to prevent counterfeiting while satisfying the regulatory requirements for a machine-readable marking in the form of a barcode or data matrix. In addition, traceability can be expanded back to the design process when the inherent markings are integrated. Moreover, this simultaneously enables inclusion of the manufacturing process into the traceability chain as well (see Figure 3). Either way, the greatest potential for innovation is seen in the ability to clearly identify the implant after surgery, preferably via noninvasive technologies already available in the hospital or technologies that can be provided without great technological effort and financial investment. Either way, the inherent feature could act as a key to access distributed information (see Figure 2) after signal processing and appropriate decoding (Figure 4).

Figure 4. Principle of inherent markings in an additively manufactured implant with part identification and tracking.



Matvieieva et al [13] performed an identifier read-out applying eddy current (EC), ultrasound (US), and micro-computed tomography (CT; also in x-ray mode) as nondestructive methods. The EC, US, and x-ray techniques allow the receipt of part of the inherent data encoded in 1D and 2D codes. The feasibility of the methods was shown for 1D Pharmacode, UDI 1D Barcode (ISO 128), and UDI 2D DataMatrix for titanium, titanium alloy, and stainless steel. Due to the physical restrictions of the chosen nondestructive methods (eg, penetration depth, magnetic and acoustic wave transmission, propagation, and density), the obtained identifiers' signals have to be processed. In fact, after the identifier´s read-out, the obtained signals are processed by

morphological and mathematical operations; once decoded, they enable further linkage with data or information stored in a database (Figure 4).

Technical Focus Data

The starting point is the construction of a 3D model of a hip implant (Figure 5) based on anthropometric data or even patient-specific information in which the coding is also integrated. Considering typical hip implant sizes, geometric complexities, and quantities, 3D printing (additive manufacturing [AM]) is increasingly developing as a competitive manufacturing method [14]. Suitable AM procedures are primarily selective laser melting (SLM) [15,16] or electron beam melting [17,18]. In addition, layer-wise build-up and achievable resolution are very beneficial for the integration of inherent features during manufacturing (see Figure 3). Using these manufacturing processes, the 3D data model needs to be converted into a facet model first (mesh) [19]. Then, the parts are positioned in the build chamber, supported, and sliced using standard AM software [20]. Based on material selection criteria such as biocompatibility, the Young modulus, strength, and fatigue strength, a (certified) raw material is selected [21-23] that is particularly tailored to the AM process requirements [24]. Material selection criteria like particle size, particle size distribution, and morphology or chemical properties are continuously checked and monitored [25,26]. The SLM process data, for example, have to be qualified for applications like implants and are subdivided into predefined parameters (considered static) and parameters to be controlled by in-process sensing (considered dynamic) [27]. Either way, essential parameters are monitored and archived [28-30]. The same applies to processing data from postheat or pressure treatment [31] as well as from destructive and nondestructive material testing [32]. Moreover, mechanical postmachining, performed to adapt the implant to the recipient needs, generates data [33]. An example of this is the description of the geometric interface to the patient (bone-implant interface) including parameters such as surface roughness. The result, however, is an extensive description of the implant and the implant creation process as well as additional information such as corresponding implant tools (Figure 6). Obviously, some of this information is of interest to surgeons (medical focus), whereas information about the implant recipient and the use of the implant (social focus) are of interest to the manufacturer (Figure 2). This means, regardless of the individual legal framework, the possibility of controlled linkage of information is seen as desirable.

Figure 5. Example of the component (part) identification using a Pharmacode integrated into the hip prosthesis and data extraction from the identifier in the implanted state by computed tomography. CAD: computer-aided design; LMS: laboratory management system.

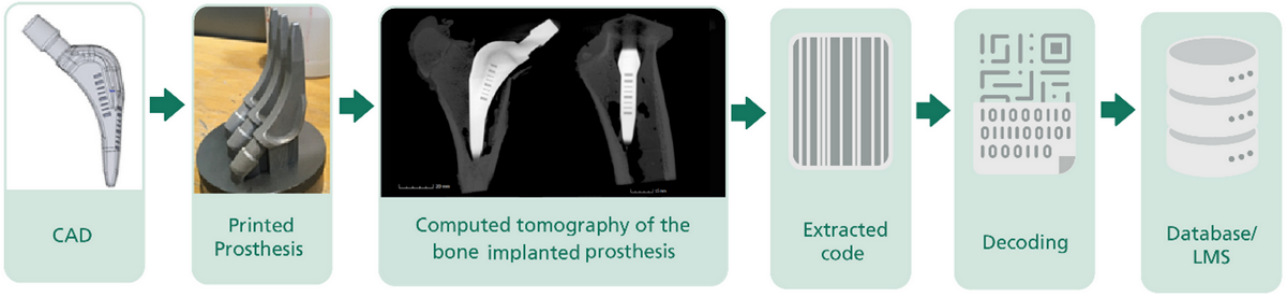
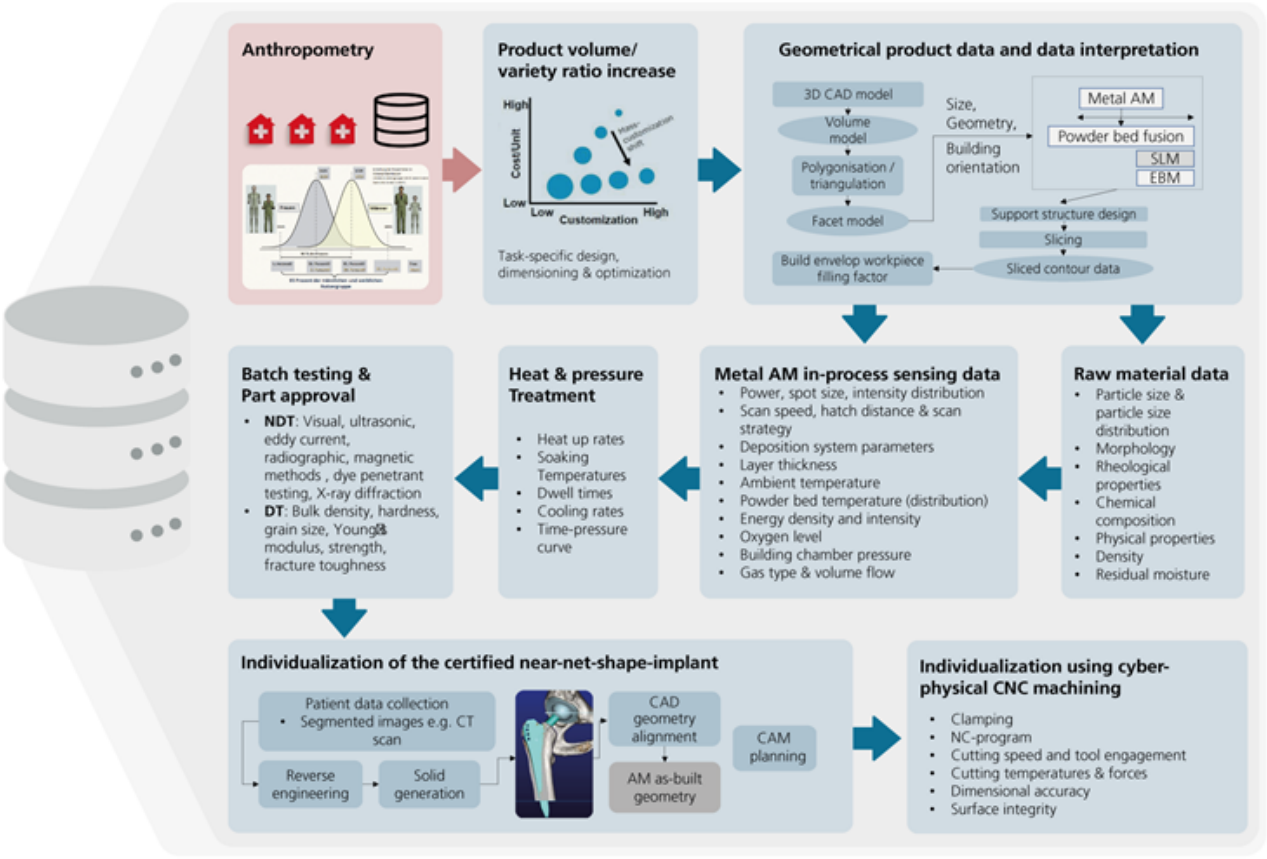


Figure 6. Schematic illustration of the process chain for additive manufacturing and subtractive individualization of hip implants with a selection of quality-determining parameters that are monitored and archived in process databases as part of quality assurance. AM: additive manufacturing; CAD: computer-aided design; CAM: computer-aided manufacturing; CNC: computer numerical control; CT: computed tomography; DT: destructive testing; EBM: electron beam melting; NC: numerical control; NDT: nondestructive testing; SLM: selective laser melting.



Social Focus Data

The social background of an implant-receiving patient (Figure 2) includes individual characteristics and conditions such as age, gender, lifestyle, and constitution type, which is of crucial importance for the formation of diseases, their duration, and treatment. The continuous monitoring or even documentation of this background can be ensured by a variety of technologies from the field of mobile health (mHealth), which allow broad mapping of dynamic data sets such as lifestyle and physical activities [34]. mHealth is an aspect of eHealth, although there is no universal definition of mHealth [1]. However, there is consensus that mHealth can be understood as medical and public health practice supported by mobile devices, such as mobile phones, patient monitoring devices, personal digital assistants, and other wireless devices [35]. This means that mHealth can be understood as "the use of mobile communications for health information and services" as patient-individual behavior without direct involvement of the health service provider [36]. Here, mHealth is seen as the technical prerequisites to monitor and document the social focus data during healing (Figure 7) or even beyond (see Figure 2).

Figure 7. Illustration of close monitoring of healing using mobile health (mHealth) applications.

In addition to the fact that mobile communication and audiovisual interaction are the decisive enablers for mHealth, there is also the aspect of powerful sensor-based hardware and high flexibility in software development for smartphones and, to a limited extent, wearables [37]. In contrast to smartphones, wearables can be designed very specifically and can therefore be used for special applications (eg, sleep monitoring) [38]. Widespread sensor systems for smartphones include [39-41] light sensor technology (ambient light, camera system in combination with lighting), proximity sensors, acceleration sensor technology, rotation sensor technology (gyroscope), electromagnetic sensor technology, digital compass (magnetometer), acoustic sensor technology via microphone, and sensor technology for location tracking (GPS).

In contrast to smartphones, wearables are more specific for a particular application and therefore have higher specificity for the integrated sensors, so that headbands can, for example, derive targeted electroencephalography [42]. This means that targeted data collection is possible with the help of sensor technology, internal processing by specific software (eg, apps), visualization based on these data, and mostly wireless communication to third parties in the form of a uniform data image. For example, the following parameters can be acquired by means of wearables and mobile devices: heart rate and pulse oxymetry readings with photoplethysmography [43] and systems to monitor activity and sleep [11].

Hence, mHealth is seen as an enabler that contributes to rehabilitation by providing valuable data about the rehabilitation measures and patient-specific activities (Figure 6) that can be stored in the EHR. In addition, realistic load scenarios can be determined that could contribute to the further optimization of hip implants (Figure 2)—to promote healing [44] while targeting shorter hospital stays and lower treatment costs, for example [45]. In addition, mHealth can help achieve a consistent database (Figure 2) to evaluate the optimal intensity, frequency, and effects of rehabilitation from a wide variety of patients over a longer period of time as these data are currently not available or insufficient [46]. Either way, linking social focus data and production data could enable significant improvements with regard to determining the actual wearing of the implant, for example, based on characteristics such as posture, weight, and movement profiles, which are summarized here as social focus data (Figure 2).

Medical Focus Data

Both routine diagnostics and revision surgery require information about implants that were placed decades ago. Specific identification using only medical imaging is currently not possible and requires access to the documentation by the initial treating physician or hospital (Figure 2). For elective as well as acute medical interventions, this documentation is not available or only available with enormous effort. With inherent markings in the implant, it is feasible to obtain information relevant for revision surgery or routine diagnostics even after

the insertion of an implant into the human body [47] (Figure 4). To assure valid and reliable surgical planning, medical data from past treatments, especially the surgical processes, follow-up examinations, and rehabilitation measures, are needed. Insufficient information about the implant and medical focus data (Figure 2) could lead to complications during revision surgery; therefore, considerable efforts are being made to obtain this information, which significantly extends the preoperative time in the hospital and results in additional costs. In fact, an example of the necessary information about the particular implant concerns the appropriate revision instruments and the existing implant components [48-50]. Moreover, the research effort prior to the operation is continuously increasing, linked to the increasing number of operations and growing variety of implant types, sizes, variants, and material combinations. In addition to knowledge on the implant system used, information on the initial implantation process, which cannot be seen in medical images (CT, x-ray), is highly relevant for a gentle and successful revision of a hip implant. In fact, insufficient information clearly increases the risk of complications for the patient. Here, it shall be emphasized that the use of unsuitable revision instruments causes the risk of an enlargement of the wound surface resulting from an invasive procedure. This, in turn, increases the risk of infection and bleeding or even periprosthetic fracture. In addition, the inevitable prolongation of the surgery time can lead to a higher anesthetic risk, increased risk of thrombosis, and unstable cardiovascular function. Consequently, there is the obligation to report “incidents” in connection with medical devices to the German Federal Institute for Drugs and Medical Devices (BfArM), which includes an indication of the cause. Hence, unique inherent identification (Figure 4) and (partial) networking of information (Figure 2) in a transparent and retraceable database seem promising to assure a higher quality of health care [11], if unauthorized access is avoided.

Results

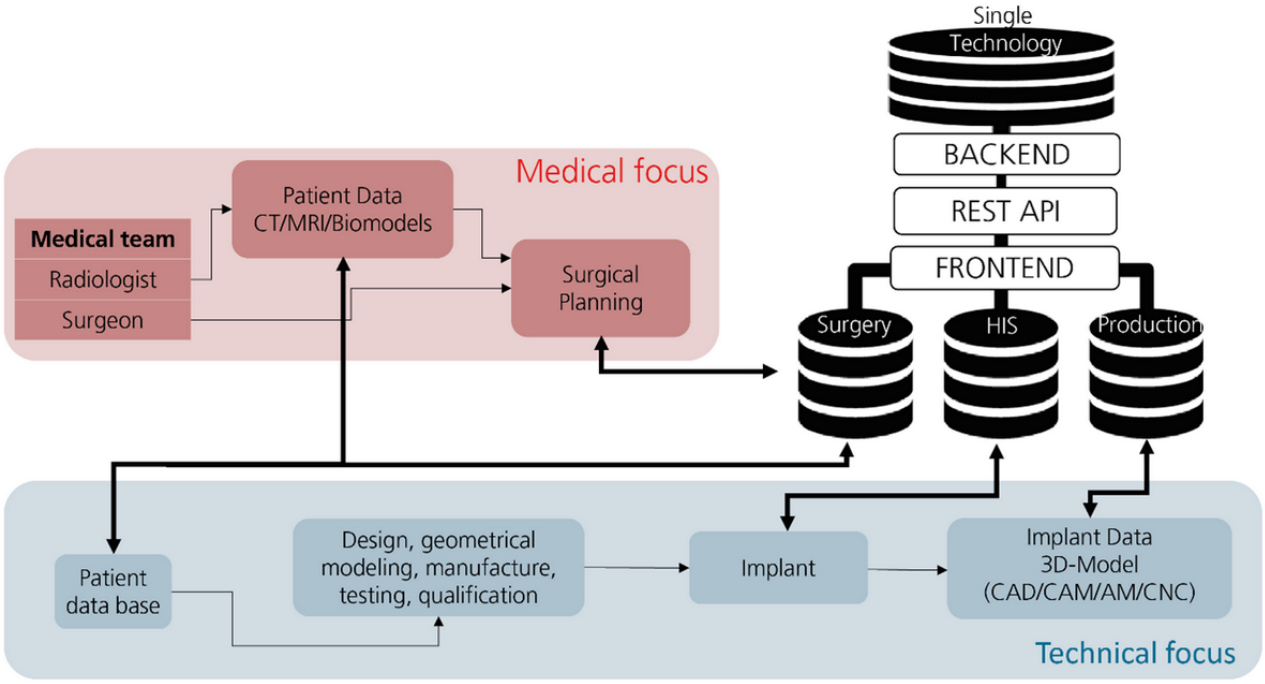
Database and Data Access

Data Integration From Heterogeneous Domains

Data integration is a crucial issue in the environment of heterogeneous patient-production data sources (Figure 2). First, there are heterogeneous data types and formats located in different databases, which implies that solving data integration challenges is a prerequisite for gaining useful information and knowledge based on appropriate analytical methods. Although concepts for databases to process heterogeneous data sets exist [51], the Laboratory Management System 4.0 (LMS 4.0) was developed specifically for the purpose of taking into account nonmedical stakeholders.

Exemplary implementation and application of this database structure were successfully demonstrated for the case of operations on the lip-jaw-palate region [52]. LMS 4.0 enables requesting data from different locations (eg, surgery, the HIS, or an implant producer) as a routine using web user interfaces. Using LMS 4.0, the surgeon collects magnetic resonance imaging results, for example, from HIS, checks patient data stored there, and has access to the integrated technical focus data (Figure 8) while planning the operational approach.

Figure 8. Partial networking of distributed data and information from different domains. AM: additive manufacturing; API: application programming interface; CAD: computer-aided design; CAM: computer-aided manufacturing; CNC; computer numerical control; CT: computed tomography; HIS: hospital information system; MRI: magnetic resonance imaging.



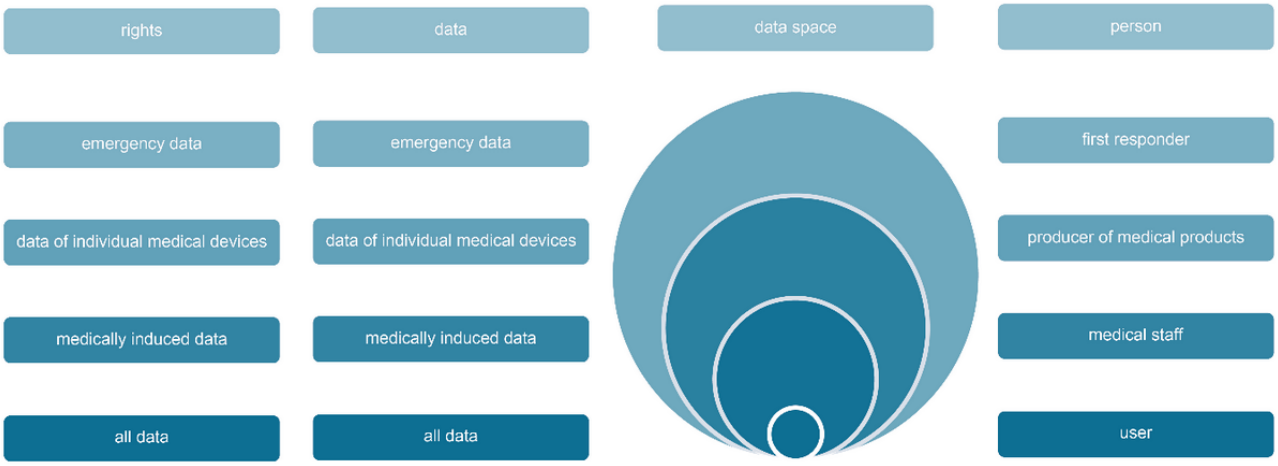
Based on integrated patient and production data, which can be expanded to an implant database that covers several manufacturers, surgical planning is simplified in order to determine which bone implant is the most suitable in the particular case, for example. Moreover, LMS 4.0 generates a dashboard and report that help the operating staff prepare for

the surgery with the selected implants through the automatic output of the associated tools and instruments. Moreover, the surgeon can use this report to prove his or her preparations for the procedure and comprehensively explain the operation to the patient. LMS 4.0 presents an architecture that implements data integration in the hospital from the production, surgery preparation, and patient data. LMS 4.0 integrates databases without any changes to the individual databases (SQL database, software back end, application programming interfaces [APIs], front end) nor any need to maintain another database. The solution combines database technology and a wrapper layer known from extract, transform, load (ETL) systems and brings it to the SQL database, WEB API (back end) layer, interface layer (REST API), and front end. It also provides semantic integration through a connection mechanism between data elements. The solution allows for integration of patient, surgery, and production data in one technological framework: data management platform and implementation of analytical methods in one end user environment. The patient data (see Figure 2) are transferred, securely, to a HIS. Medical data storage in LMS 4.0 offers a highly scalable clinic web storage service that uses cumulative digital objects (eg. patient, surgery, implant) rather than blocks or files. Object storage typically stores data, along with metadata that identify and describe the content. For metadata management and automated quality control and data fusion (ETL processes), a data consistency model (LMS I4.0 metamodel) is used to enable eventual consistency for updates or deletes to existing objects.

Role Concept for Secure Data Access

On the basis of the information domains and the links and interfaces shown in Figure 2, it becomes obvious that the database structure and the information technology (IT) system design in the back end (see Figure 8) have to accommodate different user roles to protect secure data access to sensitive patient data. Consequently, a role model was developed that takes into account both different users or user groups (eg, patients, medical staff, manufacturers of medical products, and first aid providers) as well as special situations (eg, emergency access). Deviating from static access models (role-based access control) as well as traditional shell models in this case, pure login information was linked with additional contextual information (attributed-based access control) in order to allow hierarchical access control. Either way, the principle is shown in Figure 9, indicating the 4 basic roles embedded in the defined shell model.

Figure 9. Principle of hierarchical data access based on a shell embedded role model.



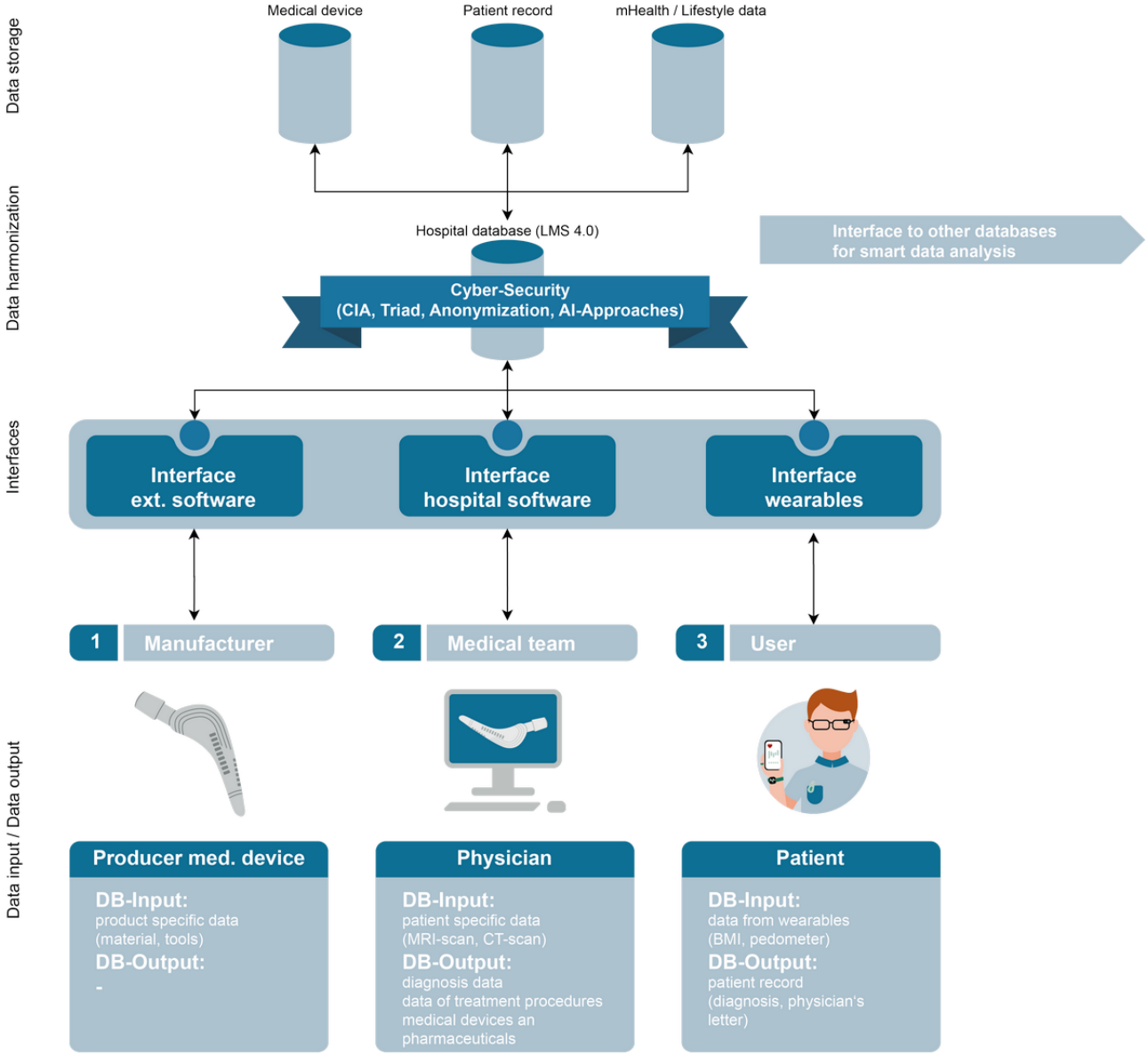
Nevertheless, the person category is subdivided into 4 roles: “first responder,” “producer of medical products,” “medical staff,” and “user.” Users can view all data via a terminal device after registration, for example, with a digital health card equipped with a radio-frequency identification transponder (radio-frequency identification), and allowed to only make entries in a dedicated area of the social focus data section (Figure 2). The write permissions include data integration from fitness watches, training performance in rehab or sports facilities, and self-collected nutritional data, while interfaces to cell phone apps, for example, are available to substitute manual input. On the other hand, the medical staff can read medically relevant data and has the right to make entries in the medical focus data (Figure 2), which shows who made the entry. This corresponds to an entry in the EHR stored in the HIS (Figure 7). Producers of medical products only have access to the medical device data area, summarized as production (Figure 7). Technical focus data (Figure 2; eg, appropriate revision instruments as described in the Medical Focus Data section) are stored here, while the patient and implant are linked in the medical focus database (Figures 2 and 7). This information is requested from the manufacturer via a modified procurement process, which ensures that the agreed data are available before the invoice for the implant is paid. Consequently, all relevant product information is stored, enabling the simplification of follow-up treatments, support for minimally invasive interventions, and excluding of medical interaction. The information transfer to the manufacturer (see the Technical Focus Data section), on the other hand, can be enabled using a data integration center with a data use and access committee for research inquiries, as is currently being developed by Prokosch et al [53]. The last role in the person category is the first responder (Figure 8), which is introduced to explain the dynamic access approach. The first responder occurs in case of an accident or emergency when lifesaving measures, for example, are necessary. For example, access is granted for a certain period if several predefined factors that were detected using a fitness watch or other smart device take effect at the same time (eg, oxygen saturation in the blood, blood

pressure, or other health-endangering characteristics). However, these attributes are securely transferred to the LAB 4.0 database management system (see the Data Integration From Heterogeneous Domains section) to obtain the necessary information depending on the authorization or to allow the addition of data. For this purpose, a standardized, well-defined interface is used to realize the data exchange and integrate smart devices for pure information retrieval as well as to develop software extensions that can be used to store the data in the database while complying with access restrictions. This enables the creation of a digital ecosystem for different participants to provide patients with optimal and, above all, digital, end-to-end health care while providing adaptive access regulations meeting authenticity requirements and assuring authenticity and appropriate access tracking. A method of secure patient-centered management of EHR data, though it can also be further processed in a deidentified format for statistical purposes, has been demonstrated with blockchain technology using cancer care as an example [54].

Data Integration Scenario

Using the hospital database LMS 4.0 (Figure 8), all individual elements of the data ecosystem are presented and explained in reference to Figure 10. In fact, the data ecosystem is divided into the following 4 levels, which are distinguished for functional structuring: “data storage,” “data harmonization,” ”interfaces,“ and ”data input/data output.“ The ”data storage“ level contains different relational databases, which, again, contain medically relevant data (medical device [see the Technical Focus Data section], patient record [see the Medical Focus Data section], and health [see the Social Focus Data section]). Data preprocessing is performed at the ”data harmonization“ level (see Figures 8 and 10), which means that incoming data are adapted to the requirements of the LMS 4.0 database structure (Figure 8) and sorted, while outgoing (anonymized) data (eg, via the data integration center) are transferred via defined data exchange procedures with strictly recorded accesses. In the underlying, but closely related, ”interfaces“ level (Figure 10), interfaces are established by extensible middleware to communicate with the ”hospital database“ (see Figure 8). This enables the integration of data users from different domains and querying data from the database. The 3 levels ”interfaces,“ ”data harmonization,“ and ”data storage“ are subjected to the CIA (confidentiality, integrity, availability) triad and ensure the functionality of the system. The ”data input/data output“ level connects the ”hospital database“ with the environment. For example, manufacturers of medical devices can transfer product-specific data into the database to make the data available to hospital staff (see Figures 3 and 10). Likewise, patients can store their vital signs from wearables, for example, in this database to support long-term examinations or enable access in medical emergencies through attribute-based access control (see the Role Concept for Secure Data Access section). At the same time, patients can see their EHR, read digital doctor's notes, or view exam results. Physicians have interfaces to both connect medical exam machines to the database and write data; medical staff can also store information, and the data can be viewed hospital-wide and processed with appropriate IT systems.

Figure 10. Data integration ecosystem with hierarchical data access. AI: artificial intelligence; CIA: confidentiality, integrity, availability; CT: computed tomography; DB: database; ext: external; LMS: laboratory management system; med: medical; mHealth: mobile health; MRI: magnetic resonance imaging.



Discussion

Principal Findings

The connection of different data lakes, beginning with implant design including the entire manufacturing to the medical treatment process as well as the tracking of lifetime characteristics inside of a data integration ecosystem, shown in this paper can disrupt today's health care system, leading to a cost-efficient, personalized system. The strict hierarchical data access concept based on a shell-embedded role model can be used to handle highly sensitive data and as a template to help clarify legal issues.

The overriding goal is to use digitization to improve the networking of interdisciplinary domains and to create secure interfaces for exchange as a prerequisite for intelligent data analysis. For this purpose, a representative hip implant application scenario was chosen due to an existing network of social, medical, and technical domains. Moreover, the interaction resulted from a skill- and experience-based union of implant and recipient results in an individual constellation that is subject to change over time. A key element in the resulting constellation is the UDI based on an inherent feature, which can be read out noninvasively after implantation. The permanently readable feature acts as a key to technical focus data, which represent testable or documented properties that are made available by the provider and are of direct or downstream interest. Consequently, we showed how the technical focus data can be integrated into existing data ecosystems. This, however, was only approached at the hospital level, which is explained by the unclear legal framework and the missing data infrastructure for a broader context. Nevertheless, the consolidation of distributed databases in a single technology solution is a scalable concept that can be transferred from a single hospital to a global solution. Another important aspect is that the introduced hierarchical data access is based on a shell-embedded role model and staggered user rights. Here, the attribute-based access control shall be emphasized because this represents nonrigid boundary

conditions in preparation for future regulations. The selected user profiles and the granted rights, on the other hand, are only examples that are up for discussion and need to be specified and challenged in further research. However, the data integration scenario distinguishes 4 levels of action layering data storage, data harmonization, interfaces, and the data input/data output layer, which harmonizes the application scenario and digital ecosystem. Nevertheless, future research must show how real benefit can be created through data linkage and how this can be monetized. Balancing the personal rights of the individual while achieving sustainable technological innovation is seen as the central challenge, which must be faced in a global context.

Conclusions

Personalized medicine requires cross-domain linkage of data, which, in turn, requires an appropriate data infrastructure and adequate hierarchical data access solutions.

Hip implant is a prime example of the usefulness of cross-domain linkage of data because it bundles social factors of the individual patient, medical aspects in the context of the implantation, and technical aspects of the implant.

UDI in terms of inherent identifiers can be the key to (selective) long-term data access especially if the postoperative readout is guaranteed.

SLM and electron beam melting make it possible to integrate features already inherent in the design process to close the traceability gap.

It is necessary to open existing databases using suitable interfaces for secure integration of data from end devices (eg, wearables or end users) and to assure availability through suitable access models (role-based, attribute-based, hybrid) while guaranteeing long-term independent data persistence.

A suitable strategy requires the combination of technical solutions from the areas of data storage, cryptographic procedures, and software engineering as well as organizational changes among the actors involved (eg, hospital staff, implant manufacturers, patients).

Holistic approaches require interdisciplinary cooperation and cross-domain data spaces, while innovative approaches and services must be developed prior or parallel to the ongoing clarification of the legal framework conditions.

To provide viable and transferable solutions at the time of legal clarification, cross-domain lighthouse projects are needed to assure the timely availability of digital business models, suitable data alliances, and an adequate digital infrastructure.

Acknowledgments

“Unique Device Identification Based Linkage of Hierarchically Accessible Data Domains in Prospective Surgical Hospital Data Ecosystems” is part of the Fraunhofer Light-house Project “futureAM - Next Generation Additive Manufacturing” funded internally by the Fraunhofer-Gesellschaft e.V. Partial sponsorship was received from the Else Kröner Fresenius Center for Digital Health of the TU Dresden. The article processing charge was funded by Fraunhofer.

Authors' Contributions

KK, AS, and UT conceptualized the study. AS designed the methodology and wrote the initial manuscript draft. KK and GL developed the software. KK, NM, CN, and GL performed the investigation. SI and WGD provided the resources. KK, AS, NM, CN, UT, GL, ABA, SI, and WGD reviewed and edited the manuscript. AS, UT, KK, NM, CN, ABA, and GL created the visualizations. AS and SI supervised the study. KK and AS provided project administration. SI and WGD acquired the funding. All authors read and agreed to the published version of the manuscript.

Conflicts of Interest

None declared.

References

1. Gross TP, Crowley J. Unique device identification in the service of public health. N Engl J Med 2012 Oct 25;367(17):1583-1585. [doi: 10.1056/NEJMp1113608] [Medline: 23013051]
2. Morlock MM, Bishop N, Zustin J, Hahn M, Rüther W, Amling M. Modes of implant failure after hip resurfacing: morphological and wear analysis of 267 retrieval specimens. J Bone Joint Surg Am 2008 Aug;90 Suppl 3:89-95. [doi: 10.2106/JBJS.H.00621] [Medline: 18676942]
3. Anagnostakos K, Schmid NV, Kelm J, Grün U, Jung J. Classification of hip joint infections. Int J Med Sci 2009 Sep 01;6(5):227-233 [FREE Full text] [doi: 10.7150/ijms.6.227] [Medline: 19841729]
4. Bhavnani SP, Narula J, Sengupta PP. Mobile technology and the digitization of healthcare. Eur Heart J 2016 May 07;37(18):1428-1438 [FREE Full text] [doi: 10.1093/eurheartj/ehv770] [Medline: 26873093]
5. Vayena E, Haeusermann T, Adjekum A, Blasimme A. Digital health: meeting the ethical and policy challenges. Swiss Med Wkly 2018;148:w14571 [FREE Full text] [doi: 10.4414/smw.2018.14571] [Medline: 29376547]
6. Granja C, Janssen W, Johansen MA. Factors determining the success and failure of eHealth interventions: systematic review of the literature. J Med Internet Res 2018 May 01;20(5):e10235 [FREE Full text] [doi: 10.2196/10235] [Medline: 29716883]

7. Topol E. The creative destruction of medicine: How the digital revolution will create better health care. New York, NY: Basic Books; 2013.
8. Hansen A, Herrmann M, Ehlers JP, Mondritzki T, Hensel KO, Truebel H, et al. Perception of the progressing digitization and transformation of the German health care system among experts and the public: mixed methods study. JMIR Public Health Surveill 2019 Oct 28;5(4):e14689 [FREE Full text] [doi: 10.2196/14689] [Medline: 31661082]
9. Hahn H, Schreiber A. E-Health. In: Neugebauer R, editor. Digitalisierung. Berlin, Heidelberg: Springer-Verlag GmbH; 2018:978-973.
10. Neugebauer R, editor. Digitalisierung. Berlin, Heidelberg: Springer-Verlag GmbH; 2018.
11. Majumder S, Mondal T, Deen MJ. Wearable sensors for remote health monitoring. Sensors (Basel) 2017 Jan 12;17(1):1 [FREE Full text] [doi: 10.3390/s17010130] [Medline: 28085085]
12. van Geerenstein D, Paul H, Zimmermann S. VDMA Studie Produktpiraterie 2018. FBA. 2018. URL: https://exportberatung.de/vdma-studie-2018-produktpiraterie/ [accessed 2022-12-16]
13. Matvieieva N, Neupetsch C, Oettel M, Makdani V, Drossel WG. A novel approach for increasing the traceability of 3D printed medical products. Current Directions in Biomedical Engineering 2020;6(3):1-4. [doi: 10.1515/cdbme-2020-3081]
14. Wohlers report 2019: 3D printing and additive manufacturing state of the industry. Wohlers Associates. 2019. URL: https://wohlersassociates.com/product/wohlers-report-2019/ [accessed 2022-12-16]
15. Abate KM, Nazir A, Jeng J. Design, optimization, and selective laser melting of vin tiles cellular structure-based hip implant. Int J Adv Manuf Technol 2021 Jan 06;112(7-8):2037-2050. [doi: 10.1007/s00170-020-06323-5]
16. Bartolomeu F, Costa M, Gomes J, Alves N, Abreu C, Silva F, et al. Implant surface design for improved implant stability – A study on Ti6Al4V dense and cellular structures produced by Selective Laser Melting. Tribology International 2019 Jan;129:272-282. [doi: 10.1016/j.triboint.2018.08.012]
17. Murr LE, Gaytan SM, Martinez E, Medina F, Wicker RB. Next generation orthopaedic implants by additive manufacturing using electron beam melting. Int J Biomater 2012;2012:245727 [FREE Full text] [doi: 10.1155/2012/245727] [Medline: 22956957]
18. Tang HP, Yang K, Jia L, He WW, Yang L, Zhang XZ. Tantalum bone implants printed by selective electron beam manufacturing (SEBM) and their clinical applications. JOM 2020 Jan 22;72(3):1016-1021. [doi: 10.1007/s11837-020-04016-8]
19. George P, Hecht F, Saltel E. Fully automatic mesh generator for 3D domains of any shape. IMPACT of Computing in Science and Engineering 1990 Sep;2(3):187-218. [doi: 10.1016/0899-8248(90)90012-y]
20. Calignano F, Galati M, Iuliano L. A metal powder bed fusion process in industry: qualification considerations. Machines 2019 Nov 13;7(4):72. [doi: 10.3390/machines7040072]
21. Gotman I. Characteristics of metals used in implants. J Endourol 1997 Dec;11(6):383-389. [doi: 10.1089/end.1997.11.383] [Medline: 9440845]
22. Heary RF, Parvathreddy N, Sampath S, Agarwal N. Elastic modulus in the selection of interbody implants. J Spine Surg 2017 Jun;3(2):163-167 [FREE Full text] [doi: 10.21037/jss.2017.05.01] [Medline: 28744496]
23. Pandey A, Awasthi A, Saxena KK. Metallic implants with properties and latest production techniques: a review. Advances in Materials and Processing Technologies 2020 Mar 09;6(2):405-440. [doi: 10.1080/2374068x.2020.1731236]
24. Tassa O, Alleva L, Sorci R. Powders for Additive Manufacturing: From Design to Certification. In: Ionescu M, Sommitsch C, Poletti C, Kozeschnik E, Chandra T, editors. Materials Science Forum. Bäch SZ, Switzerland: Trans Tech Publications Ltd; 2021:1473-1478.
25. Slotwinski JA, Garboczi EJ, Stutzman PE, Ferraris CF, Watson SS, Peltz MA. Characterization of metal powders used for additive manufacturing. J Res Natl Inst Stand Technol 2014;119:460-493 [FREE Full text] [doi: 10.6028/jres.119.018] [Medline: 26601040]
26. Harun W, Manam N, Kamariah M, Sharif S, Zulkifly A, Ahmad I, et al. A review of powdered additive manufacturing techniques for Ti-6al-4v biomedical applications. Powder Technology 2018 May;331:74-97. [doi: 10.1016/j.powtec.2018.03.010]
27. Spears TG, Gold SA. In-process sensing in selective laser melting (SLM) additive manufacturing. Integr Mater Manuf Innov 2016 Feb 11;5(1):16-40. [doi: 10.1186/s40192-016-0045-4]
28. Gaikwad A, Yavari R, Montazeri M, Cole K, Bian L, Rao P. Toward the digital twin of additive manufacturing: Integrating thermal simulations, sensing, and analytics to detect process faults. IISE Transactions 2020 Jan 24;52(11):1204-1217. [doi: 10.1080/24725854.2019.1701753]
29. Mukherjee T, DebRoy T. A digital twin for rapid qualification of 3D printed metallic components. Applied Materials Today 2019 Mar;14:59-65. [doi: 10.1016/j.apmt.2018.11.003]
30. Liu C, Le Roux L, Körner C, Tabaste O, Lacan F, Bigot S. Digital twin-enabled collaborative data management for metal additive manufacturing systems. Journal of Manufacturing Systems 2022 Jan;62:857-874. [doi: 10.1016/j.jmsy.2020.05.010]
31. Rudskoy AI, Kolbasnikov NG. Digital twins of processes of thermomechanical treatment of steel. Met Sci Heat Treat 2020 Jun 20;62(1-2):3-10. [doi: 10.1007/s11041-020-00505-4]
32. Seifi M, Salem A, Beuth J, Harrysson O, Lewandowski JJ. Overview of materials qualification needs for metal additive manufacturing. JOM 2016 Jan 27;68(3):747-764. [doi: 10.1007/s11837-015-1810-0]

33. Seidel A, Gollee C, Schnellhardt T, Hammer M, Dassing J, Vogt R, et al. Cyber-physical approach toward semiautonomous postprocessing of additive manufactured parts and components. Journal of Laser Applications 2021 Feb;33(1):012033. [doi: 10.2351/7.0000328]
34. Al Ayubi SU, Parmanto B, Branch R, Ding D. A persuasive and social mHealth application for physical activity: a usability and feasibility study. JMIR Mhealth Uhealth 2014 May 22;2(2):e25 [FREE Full text] [doi: 10.2196/mhealth.2902] [Medline: 25099928]
35. Ryu S. Book review: mHealth: new horizons for health through mobile technologies: based on the findings of the second global survey on eHealth. Healthcare Informatics Research 2012;18(3):231. [doi: 10.4258/hir.2012.18.3.231]
36. Nacinovich M. Defining mHealth. Journal of Communication in Healthcare 2013 Jul 18;4(1):1-3 [FREE Full text] [doi: 10.1179/175380611X12950033990296]
37. Iso-Ketola P, Karinsalo T, Vanhala J. HipGuard: A wearable measurement system for patients recovering from a hip operation. 2008 Presented at: Second International Conference on Pervasive Computing Technologies for Healthcare; January 30 - February 1, 2008; Tampere, Finland. [doi: 10.4108/icst.pervasivehealth2008.2580]
38. Guillodo E, Lemey C, Simonnet M, Walter M, Baca-García E, Masetti V, HUGOPSY Network, et al. Clinical applications of mobile health wearable-based sleep monitoring: systematic review. JMIR Mhealth Uhealth 2020 Apr 01;8(4):e10733 [FREE Full text] [doi: 10.2196/10733] [Medline: 32234707]
39. Freidlin RZ, Dave AD, Espey BG, Stanley ST, Garmendia MA, Pursley R, et al. Measuring risky driving behavior using an mHealth smartphone app: development and evaluation of gForce. JMIR Mhealth Uhealth 2018 Apr 19;6(4):e69 [FREE Full text] [doi: 10.2196/mhealth.9290] [Medline: 29674309]
40. Grossi M. A sensor-centric survey on the development of smartphone measurement and sensing systems. Measurement 2019 Mar;135:572-592. [doi: 10.1016/j.measurement.2018.12.014]
41. Majumder S, Deen MJ. Smartphone sensors for health monitoring and diagnosis. Sensors (Basel) 2019 May 09;19(9):1 [FREE Full text] [doi: 10.3390/s19092164] [Medline: 31075985]
42. Piwek L, Ellis DA, Andrews S, Joinson A. The rise of consumer health wearables: promises and barriers. PLoS Med 2016 Feb;13(2):e1001953 [FREE Full text] [doi: 10.1371/journal.pmed.1001953] [Medline: 26836780]
43. Castaneda D, Esparza A, Ghamari M, Soltanpur C, Nazeran H. A review on wearable photoplethysmography sensors and their potential future applications in health care. Int J Biosens Bioelectron 2018;4(4):195-202 [FREE Full text] [doi: 10.15406/ijbsbe.2018.04.00125] [Medline: 30906922]
44. Shah FA, Snis A, Matic A, Thomsen P, Palmquist A. 3D printed Ti6Al4V implant surface promotes bone maturation and retains a higher density of less aged osteocytes at the bone-implant interface. Acta Biomater 2016 Jan;30:357-367. [doi: 10.1016/j.actbio.2015.11.013] [Medline: 26577985]
45. Munin MC, Rudy TE, Glynn NW, Crossett LS, Rubash HE. Early inpatient rehabilitation after elective hip and knee arthroplasty. JAMA 1998 Mar 18;279(11):847-852. [doi: 10.1001/jama.279.11.847] [Medline: 9515999]
46. Khan F, Ng L, Gonzalez S, Hale T, Turner-Stokes L. Multidisciplinary rehabilitation programmes following joint replacement at the hip and knee in chronic arthropathy. Cochrane Database Syst Rev 2008 Apr 16;2008(2):CD004957 [FREE Full text] [doi: 10.1002/14651858.CD004957.pub3] [Medline: 18425906]
47. Marschner U, Grätz H, Jettkant B, Ruwisch D, Woldt G, Fischer W, et al. Integration of a wireless lock-in measurement of hip prosthesis vibrations for loosening detection. Sensors and Actuators A: Physical 2009 Nov;156(1):145-154. [doi: 10.1016/j.sna.2009.08.025]
48. Volkmann R. [Hip implant revision. Avoiding mistakes and managing risk]. Orthopade 2009 Aug;38(8):718-728. [doi: 10.1007/s00132-009-1427-5] [Medline: 19672577]
49. Imhoff AB, editor. Schulter/Ellenbogen/Stoßwelle/Hüfte. Berlin, Heidelberg: Steinkopff Heidelberg; 1999.
50. Nogler M, Thaler M. [Surgical access routes to the hip joint in the elderly]. Orthopade 2017 Jan;46(1):18-24. [doi: 10.1007/s00132-016-3366-2] [Medline: 28004127]
51. Brinker TJ, Rudolph S, Richter D, von Kalle C. Patient-centered mobile health data management solution for the German health care system (The DataBox Project). JMIR Cancer 2018 May 11;4(1):e10160 [FREE Full text] [doi: 10.2196/10160] [Medline: 29752255]
52. Schröder TA, Maiwald M, Reinicke A, Teicher U, Seidel A, Schmidt T, et al. A holistic approach for the identification of success factors in secondary cleft osteoplasty. J Pers Med 2022 Mar 21;12(3):1 [FREE Full text] [doi: 10.3390/jpm12030506] [Medline: 35330506]
53. Prokosch H, Acker T, Bernarding J, Binder H, Boeker M, Boerries M, et al. MIRACUM: Medical informatics in research and care in university medicine. Methods Inf Med 2018 Jul;57(S 01):e82-e91 [FREE Full text] [doi: 10.3414/ME17-02-0025] [Medline: 30016814]
54. Dubovitskaya A, Baig F, Xu Z, Shukla R, Zambani PS, Swaminathan A, et al. ACTION-EHR: patient-centric blockchain-based electronic health record data management for cancer care. J Med Internet Res 2020 Aug 21;22(8):e13598 [FREE Full text] [doi: 10.2196/13598] [Medline: 32821064]

Abbreviations

AM: additive manufacturing
API: application programming interface
CIA: confidentiality, integrity, availability
CT: computed tomography
EC: eddy current
ETL: extract, transform, load
HIS: hospital information system
IT: information technology
LMS: Laboratory Management System
MDR: Medical Device Regulation
mHealth: mobile health
PHR: personal health record
SLM: selective laser melting
UDI: unique device identification
US: ultrasound

Edited by C Lovis; submitted 02.08.22; peer-reviewed by N Karnik; comments to author 10.11.22; revised version received 23.11.22; accepted 24.11.22; published 27.01.23.

Please cite as:
Kozak K, Seidel A, Matvieieva N, Neupetsch C, Teicher U, Lemme G, Ben Achour A, Barth M, Ihlenfeldt S, Drossel WG
Unique Device Identification–Based Linkage of Hierarchically Accessible Data Domains in Prospective Surgical Hospital Data Ecosystems: User-Centered Design Approach
JMIR Med Inform 2023;11:e41614
URL: https://medinform.jmir.org/2023/1/e41614
doi:10.2196/41614
PMID:36705946

Original Paper

ChatGPT-Generated Differential Diagnosis Lists for Complex Case–Derived Clinical Vignettes: Diagnostic Accuracy Evaluation

Takanobu Hirosawa[1*], MD, PhD; Ren Kawamura[1*], MD, PhD; Yukinori Harada[1*], MD, PhD; Kazuya Mizuta[1*], MD; Kazuki Tokumasu[2*], MD, PhD; Yuki Kaji[3], MD, MPH; Tomoharu Suzuki[4], MD; Taro Shimizu[1*], MD, MSc, MPH, MBA, PhD

[1]Department of Diagnostic and Generalist Medicine, Dokkyo Medical University, Tochigi, Japan

[2]Department of General Medicine, Okayama University Graduate School of Medicine, Dentistry and Pharmaceutical Sciences, Okayama, Japan

[3]Department of General Medicine, International University of Health and Welfare Narita Hospital, Chiba, Japan

[4]Department of Hospital Medicine, Urasoe General Hospital, Okinawa, Japan

[*]these authors contributed equally

Corresponding Author:
Takanobu Hirosawa, MD, PhD
Department of Diagnostic and Generalist Medicine
Dokkyo Medical University
880 Kitakobayashi, Mibu-cho
Shimotsuga
Tochigi, 321-0293
Japan
Phone: 81 282861111
Email: hirosawa@dokkyomed.ac.jp

Abstract

Background: The diagnostic accuracy of differential diagnoses generated by artificial intelligence chatbots, including ChatGPT models, for complex clinical vignettes derived from general internal medicine (GIM) department case reports is unknown.

Objective: This study aims to evaluate the accuracy of the differential diagnosis lists generated by both third-generation ChatGPT (ChatGPT-3.5) and fourth-generation ChatGPT (ChatGPT-4) by using case vignettes from case reports published by the Department of GIM of Dokkyo Medical University Hospital, Japan.

Methods: We searched PubMed for case reports. Upon identification, physicians selected diagnostic cases, determined the final diagnosis, and displayed them into clinical vignettes. Physicians typed the determined text with the clinical vignettes in the ChatGPT-3.5 and ChatGPT-4 prompts to generate the top 10 differential diagnoses. The ChatGPT models were not specially trained or further reinforced for this task. Three GIM physicians from other medical institutions created differential diagnosis lists by reading the same clinical vignettes. We measured the rate of correct diagnosis within the top 10 differential diagnosis lists, top 5 differential diagnosis lists, and the top diagnosis.

Results: In total, 52 case reports were analyzed. The rates of correct diagnosis by ChatGPT-4 within the top 10 differential diagnosis lists, top 5 differential diagnosis lists, and top diagnosis were 83% (43/52), 81% (42/52), and 60% (31/52), respectively. The rates of correct diagnosis by ChatGPT-3.5 within the top 10 differential diagnosis lists, top 5 differential diagnosis lists, and top diagnosis were 73% (38/52), 65% (34/52), and 42% (22/52), respectively. The rates of correct diagnosis by ChatGPT-4 were comparable to those by physicians within the top 10 (43/52, 83% vs 39/52, 75%, respectively; *P*=.47) and within the top 5 (42/52, 81% vs 35/52, 67%, respectively; *P*=.18) differential diagnosis lists and top diagnosis (31/52, 60% vs 26/52, 50%, respectively; *P*=.43) although the difference was not significant. The ChatGPT models' diagnostic accuracy did not significantly vary based on open access status or the publication date (before 2011 vs 2022).

Conclusions: This study demonstrates the potential diagnostic accuracy of differential diagnosis lists generated using ChatGPT-3.5 and ChatGPT-4 for complex clinical vignettes from case reports published by the GIM department. The rate of correct diagnoses within the top 10 and top 5 differential diagnosis lists generated by ChatGPT-4 exceeds 80%. Although derived from a limited data set of case reports from a single department, our findings highlight the potential utility of ChatGPT-4 as a supplementary tool for physicians, particularly for those affiliated with the GIM department. Further investigations should explore the diagnostic

accuracy of ChatGPT by using distinct case materials beyond its training data. Such efforts will provide a comprehensive insight into the role of artificial intelligence in enhancing clinical decision-making.

(JMIR Med Inform 2023;11:e48808) doi:10.2196/48808

KEYWORDS

artificial intelligence; AI chatbot; ChatGPT; large language models; clinical decision support; natural language processing; diagnostic excellence; language model; vignette; case study; diagnostic; accuracy; decision support; diagnosis

Introduction

Decision-Making in Health Care

In health care, accurate diagnosis plays a critical role in the effective management of patients' conditions [1]. Clinicians often rely on their expertise and various case presentations to make clinical decisions. However, the increasing complexity of cases, particularly those requiring referrals to specialized departments such as general internal medicine (GIM), and the rapid expansion of medical knowledge necessitate enhanced diagnostic support. A single-center study reported diagnostic error rates of 2% in an outpatient GIM department [2], while a systematic review found that the error rates exceeded by 10% in older adult patients [3]. Such inaccuracies underline the pressing need for tools to aid physicians in making more accurate diagnoses [4]. One promising avenue being explored is the application of clinical decision support (CDS) systems.

CDS Tools

Various CDS systems, including symptom checkers [5] and differential diagnosis generators [6], have been developed over the years. The former are generally designed for the general public, while the latter are intended for health care providers. The journey of computer-aided health care traces back to the early 1970s, marked by a strong interest in harnessing computing power to enhance care quality. Historically, CDS tools often employ multistep processes that combine logical or computational processes, probability assessments, and heuristic methods. Notably, a combination of algorithms and heuristic rules has been integral to many medical applications [7]. There is evidence of CDS tools being utilized in the outpatient department of GIM [8]. However, despite the potential of CDS systems to boost diagnostic accuracy and efficiency, they often increase clinicians' workload [9], particularly due to the need for structured input data. This remains a great barrier to their widespread adoption. In this context, artificial intelligence (AI), especially large language models, provides an alternative approach for health care support [10], particularly through the AI chatbot [11].

ChatGPT in Health Care

AI chatbots such as ChatGPT have demonstrated potential in facilitating effective communication between patients and health care providers [12] and transforming medical writing [13]. ChatGPT, developed by OpenAI, is an application of large language model based on natural language processing, known as a generative pretrained transformer (GPT) [14]. It can generate human-like responses to user prompts. With the progression from the third-generation GPT (GPT-3.5) to the fourth-generation GPT (GPT-4), the model's accuracy has improved in professional examinations [15] and multiple-choice problems across various languages [16]. Yet, AI chatbots are not exempt from limitations and risks [17,18]. These limitations encompass transparency issues [19], nonspecialized medical knowledge, outdated medical information, inherent biases, and a potential to disseminate misinformation [11]. Despite these challenges, AI systems such as ChatGPT are continually improving and hold promise as essential tools for achieving diagnostic excellence [20].

To prepare for potential clinical applications of AI chatbots, it is essential to evaluate their diagnostic accuracy, particularly for complex cases that frequently necessitate referral to specialized departments such as the GIM department. If harnessed correctly, generative AI like ChatGPT could reduce the diagnostic errors attributed to the inherent complexity of the GIM domain. This would streamline the department's workflow, enhancing patient care and outcomes. The study will reveal the potential of generative AIs, including ChatGPT as the CDS, especially in the GIM department.

Previous studies have reported that the diagnostic accuracy of the differential diagnosis lists generated by ChatGPT for clinical vignettes falls between 64% and 83% [21,22]. A clinical vignette is a concise narrative used in research to present a clinical scenario. However, these earlier studies did not focus on the materials derived from the GIM department, which is known for its diagnostically challenging cases. This gap in the literature accentuates the novelty and distinctiveness of our study. We aimed to evaluate the diagnostic accuracy of the differential diagnosis lists generated by ChatGPT, specifically using clinical vignettes derived from case reports published by the GIM department. By focusing on these GIM case reports, our research potentially offers a more rigorous appraisal of the diagnostic prowess of ChatGPT compared to preceding studies. In line with this, we expect ChatGPT-4 to provide the correct diagnosis in its differential diagnosis lists with an accuracy consistent with or within the previously reported range of 64%-83%.

Methods

Study Design

We evaluated the diagnostic accuracy of the differential diagnosis lists generated by ChatGPT-3.5 and ChatGPT-4 for clinical vignettes from case reports published by the Department of GIM. The term "differential diagnosis" refers to a list of possible conditions or diseases that could be causing a patient's symptoms and signs. It is created by considering the patient's clinical history, physical examination, and the results of any investigations, thus aiding in the diagnostic process. This study was conducted at the GIM Department (Department of

Diagnostic and Generalist Medicine) of Dokkyo Medical University Hospital, Shimotsuga, Tochigi, Japan.

Ethical Considerations

Because this study used case vignettes from published case reports, approval by the ethics committee and requirement for individual consent were not required.

Clinical Vignettes

We used clinical vignettes from case reports published by the GIM Department of Dokkyo Medical University Hospital. Clinical cases that were challenging to diagnose and typically involved a high level of complexity were often referred to the GIM department. Some of these cases were published as case reports in medical journals. To find case reports published in English from our department, we searched PubMed using the following keywords on March 20, 2023: "(Dokkyo Medical University [affil]) AND (Generalist Medicine [affil]) AND (2016/4/1:2022/12/31 [dp]) AND (Case Reports [PT])." After finding 54 case reports in PubMed, 2 experienced GIM physicians (TH and RK) checked these case reports for diagnostic or nondiagnostic cases, assessed the final diagnosis, and displayed them as clinical vignettes. Two cases were excluded because they were nondiagnostic. In total, 52 cases were included in this study. For example, consider the case reports titled "Hepatic portal venous gas after diving" [23], which is mentioned as case number 3 in Table S1 of Multimedia Appendix 1 and Table S2 of Multimedia Appendix 2. From this report, we extracted the clinical vignette from the case description section: "A 68-year-old man with diabetes and...There was no evidence of pneumatosis intestinalis." Decompression sickness was determined as the final diagnosis for this case. These case reports meet the standards required for publication in peer-reviewed journals and have been written and selected by experienced GIM physicians. Each clinical vignette included the clinical history, physical examination, and results of the investigation. The title, abstract, introduction, clinical assessment, differential diagnosis, final diagnosis, figures, legends, tables, and case reports were removed from the vignettes. The final diagnosis for each case, which had been established through the usual diagnostic processes and subsequently published in these case reports, was assessed and displayed in the form of clinical vignettes. The final diagnosis was confirmed by 2 experienced GIM physicians. Discrepancies between the 2 physicians were resolved through discussions. We also assessed the publication date and status of the included case reports as open access.

Differential Diagnosis Lists Created by Physicians

The differential diagnosis lists for each clinical vignette were independently created by 3 other GIM physicians (KT, YK, and T Suzuki) not affiliated with Dokkyo Medical University. Each clinical vignette was allocated to 1 physician, resulting in an average of 17 case descriptions being handled by each physician. They were instructed to create the top 10 differential diagnosis lists in English by reading the same clinical vignettes, without consulting other physicians or using CDS tools. It is essential to highlight that the physicians did not adhere to any specific guidelines, criteria, or protocols during this process. They operated based solely on their expertise and experience. Before creating the differential diagnosis lists, they were confirmed to be unaware of the case reports, clinical vignettes, final diagnosis, and differential diagnosis lists generated by ChatGPT-3.5 and ChatGPT-4. The physicians also remained blinded to each other's assessments. A computer-generated order table determined the sequence in which the clinical vignettes were presented.

Differential Diagnosis Lists Generated by ChatGPT

We used ChatGPT, an application of the GPT-3.5 model (March 14 version; ChatGPT-3.5, OpenAI, LLC), on March 20, 2023. We also used ChatGPT, an application of the GPT-4 model (March 23 version; ChatGPT-4, OpenAI, LLC), on April 10, 2023. Neither of the ChatGPT models were specially trained or reinforced for medical diagnoses. The physician (TH) typed the following text in the prompt: "Tell me the top 10 suspected illnesses for the following symptoms: (copy and paste each clinical vignette)." The prompt was designed to encourage the ChatGPT models to generate a list of differential diagnoses. The rationale behind selecting this particular prompt was grounded in preliminary testing. In these tests, various prompts were evaluated for their effectiveness in soliciting a comprehensive list of potential illnesses. This prompt consistently yielded reliable and inclusive differential diagnoses in our initial evaluations.

To minimize potential bias, the order in which the vignettes were presented to ChatGPT-3.5 and ChatGPT-4 was determined using a computer-generated order table. To ensure no interference from previous responses, physicians cleared the previous conversation before introducing new clinical vignettes. We used the initial answers as the top 10 differential diagnosis lists generated by ChatGPT-3.5 and ChatGPT-4.

Evaluation of Differential Diagnosis Lists

Two other GIM physicians (YH and KM) evaluated whether the final diagnosis was included in the differential diagnosis lists created by the physicians and those generated by ChatGPT models. A diagnosis was labeled "1" if it accurately and specifically identified the condition or was sufficiently close to the exact diagnosis that it would enable prompt and appropriate treatment. Conversely, a diagnosis was marked as "0" if it diverged significantly from the actual diagnosis [24]. When the final diagnosis was present, the researcher further assessed its ranking within the list. Discrepancies between the 2 evaluators were resolved through discussions. The study design is illustrated in Figure 1. Examples of a differential diagnosis list generated by ChatGPT-3.5 and ChatGPT-4 are shown in Figures 2-3 and Figures 4-5, respectively.

Figure 1. Study design.

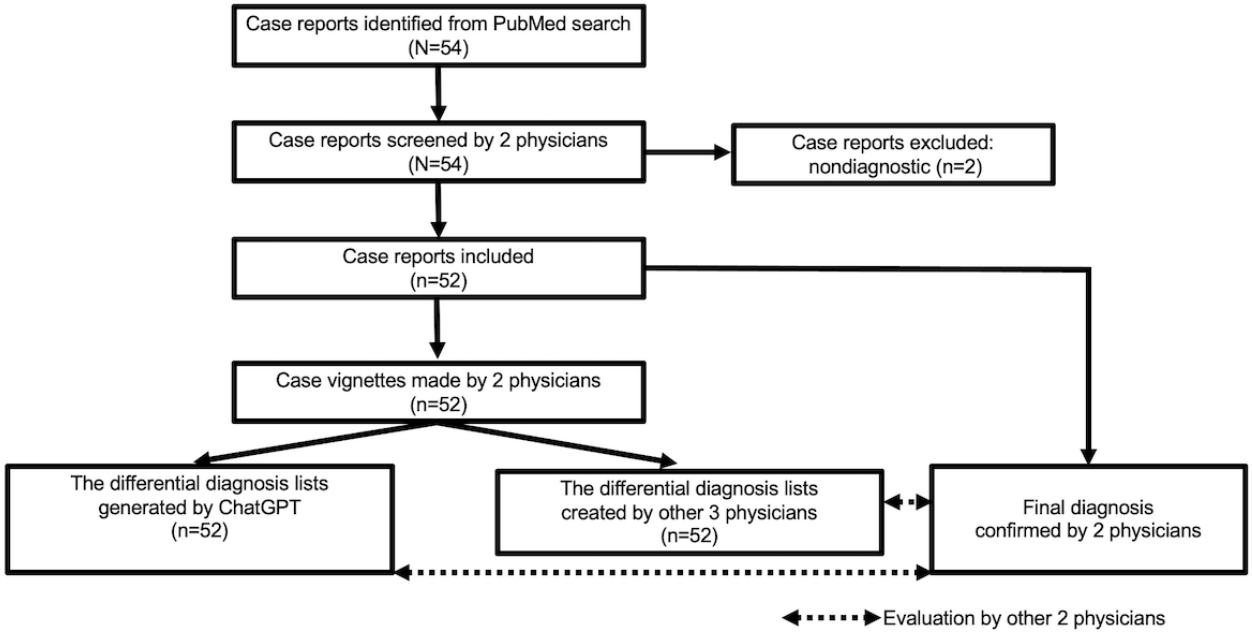


Figure 2. A differential diagnosis list generated by the third-generation ChatGPT for a sample case.

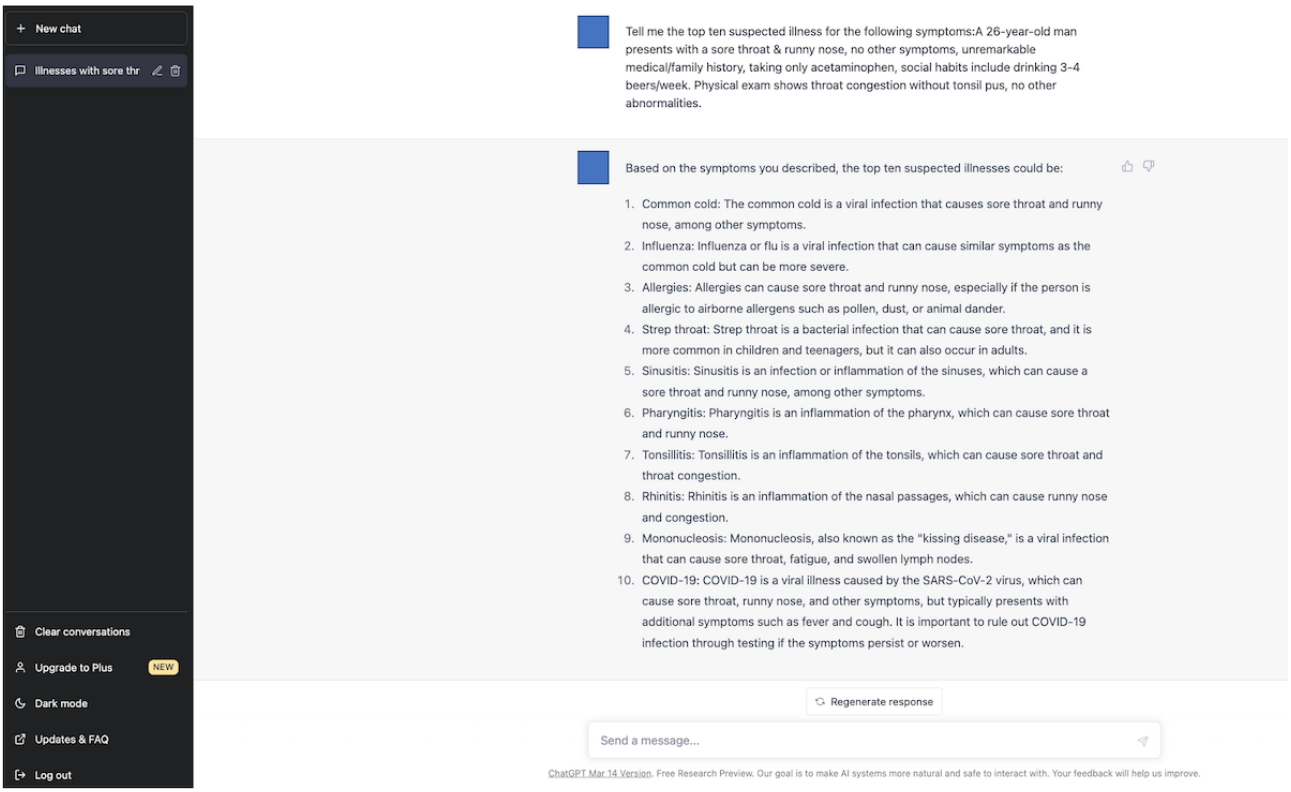


Figure 3. Explanation for the differential diagnosis list generated by the third-generation ChatGPT for a sample case. ChatGPT-3: third-generation ChatGPT.

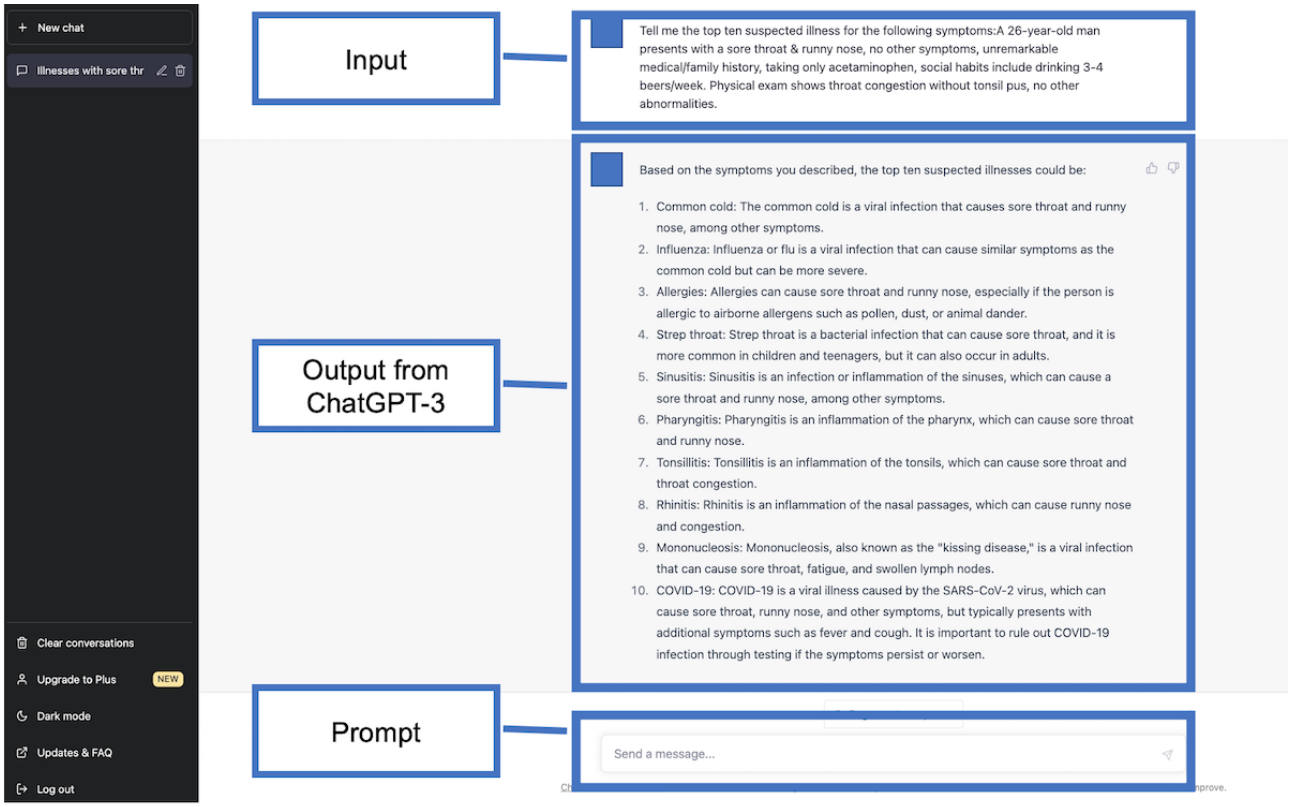


Figure 4. A differential diagnosis list generated by the fourth-generation ChatGPT for a sample case.

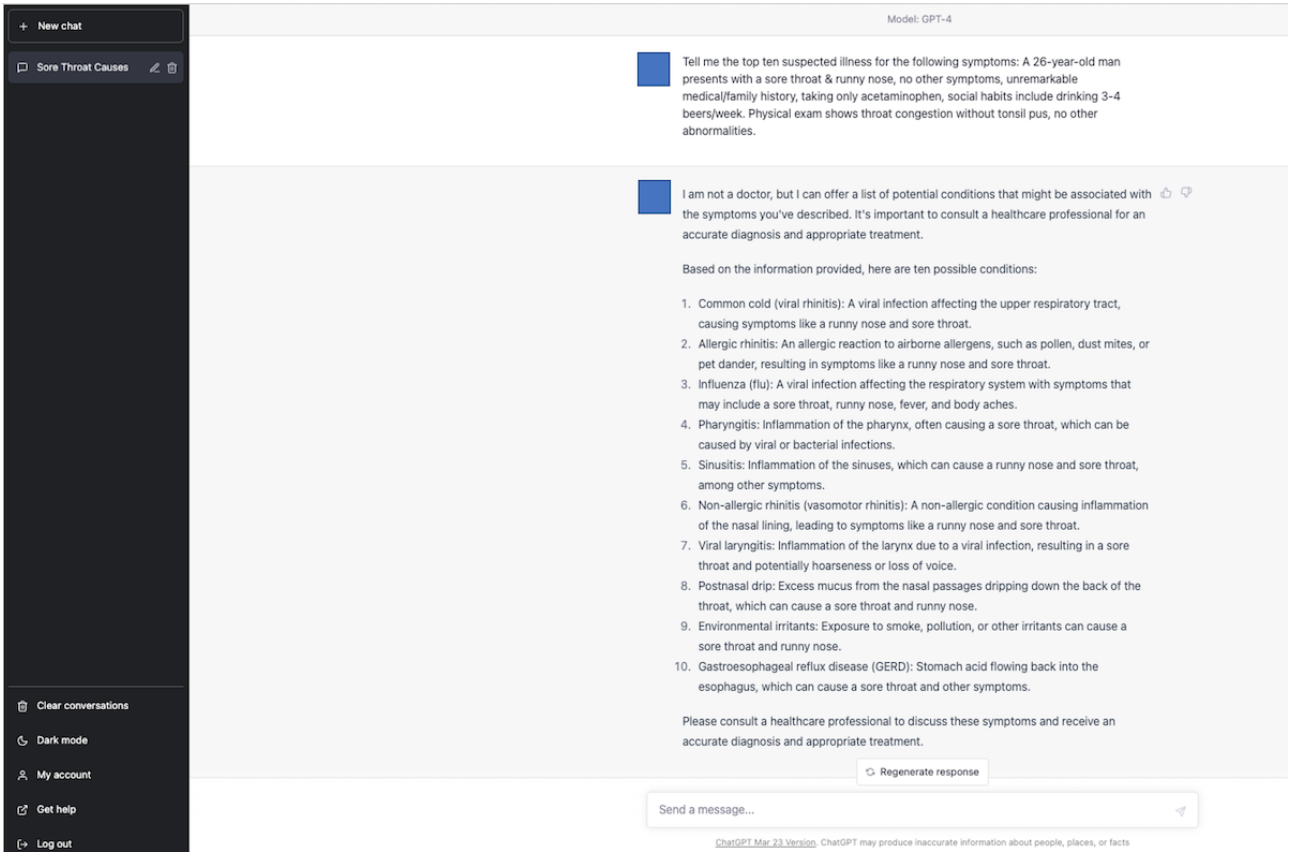
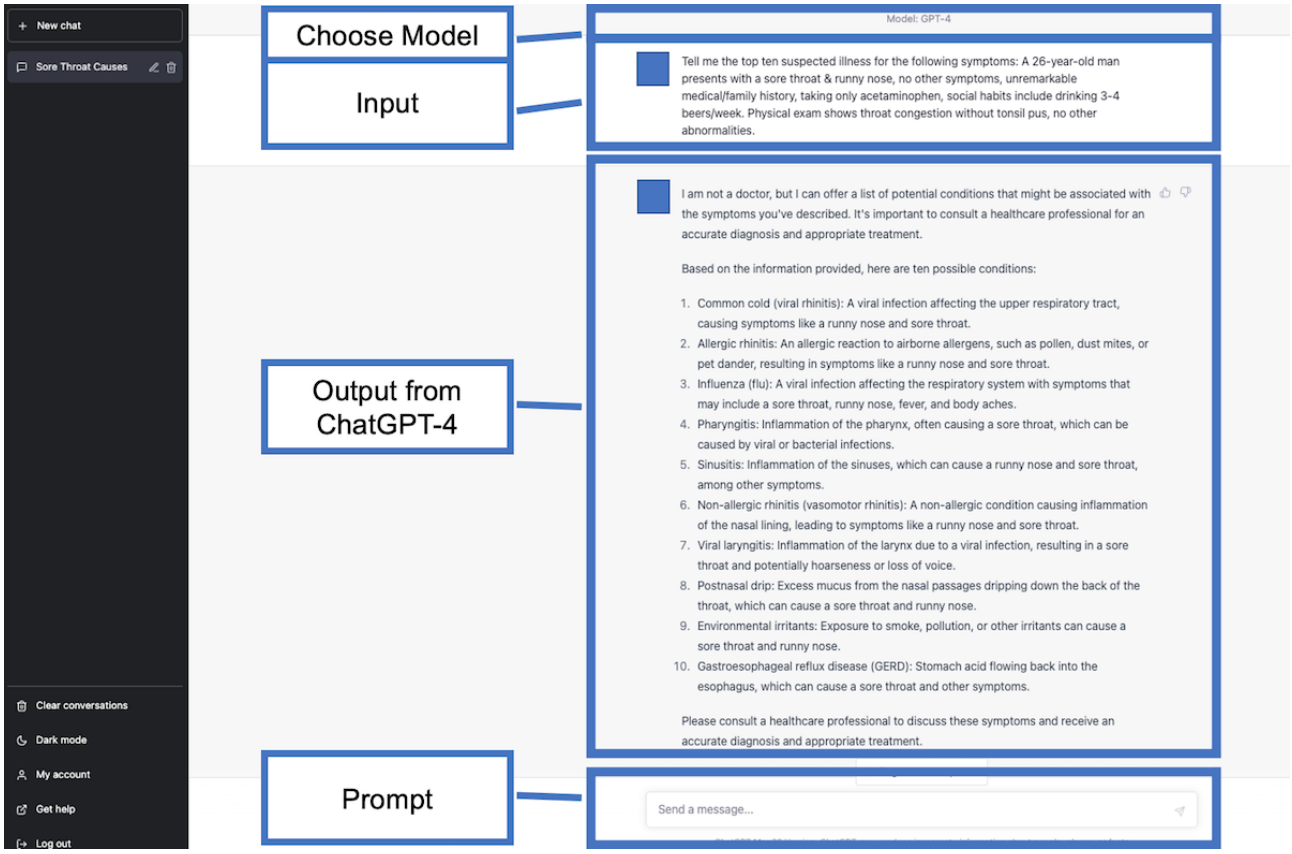


Figure 5. Explanation for the differential diagnosis list generated by the fourth-generation ChatGPT for a sample case. ChatGPT4: fourth-generation ChatGPT.



Measurements

We measured the rate of correct diagnoses within the top 10 differential diagnosis lists, top 5 differential diagnosis lists, and top diagnosis provided by ChatGPT-3.5, ChatGPT-4, and the physicians. As a binary approach, we scored the presence of the final diagnosis on the list as one and its absence as zero. For an exploratory analysis, we compared the rates of correct diagnoses in the lists generated by ChatGPT-3.5 and ChatGPT-4 between case reports that were open access and those that were not. This comparison was motivated by understanding that GPT-3.5 and GPT-4 were primarily learned from open sources available on the internet [16]. Given that these models are predominantly trained on openly accessible data, we postulated that open access case reports might yield better diagnostic results than non–open access ones. Additionally, we compared the rates of correct diagnoses within the lists generated by ChatGPT-3.5 and ChatGPT-4 based on the publishing year prior to 2021 or in 2022. This distinction arises from the knowledge cutoffs for ChatGPT-3.5 and ChatGPT-4, which were set in early 2021. Since the models would be more familiar with data before this time and less informed about subsequent publications, we hypothesized that the case reports published in the years prior to 2021 could produce better diagnostic results than those published in 2022. However, the details of the learning data source and cutoff timing were not available to the public.

Analysis

Categorical or binary variables were presented as numbers (percentages) and compared using the chi-square test. To mitigate the increased risk of type I error arising from multiple comparisons, we employed the Bonferroni correction [25]. Although alternative methods exist, we chose the Bonferroni correction for its strict control over false positives. When conducting multiple comparisons, we set the Bonferroni-corrected significance level at a P value <.02. This was derived by dividing .05 (the standard level of significance) by 3 (the number of comparisons undertaken). Both the chi-square test and the computation of the Bonferroni-corrected significance level were conducted in R (version 4.2.2; R Foundation for Statistical Computing) using the stats library (version 4.2.2).

Results

Case Report Profiles

In total, 52 case reports were included in this study, among which 39 (75%) were open access case reports. A total of 24 (46%) case reports were published prior to 2021. Of the total case reports, 12 (23%) were published in 2021 and 16 (31%) were published in 2022. The included case reports are presented in Multimedia Appendix 1.

Diagnostic Performance

Representative examples of differential diagnosis lists with the final diagnosis are shown in Table 1.

The rates of correct diagnosis by ChatGPT-4 within the top 10 differential diagnoses, top 5 differential diagnoses, and top diagnosis were 83% (43/52), 81% (42/52), and 60% (31/52), respectively (Table 2). The rates of correct diagnosis by

ChatGPT-3.5 within the top 10 differential diagnoses, top 5 differential diagnoses, and top diagnosis were 73% (38/52), 65% (34/52), and 42% (22/52), respectively. The rates of correct diagnosis by ChatGPT-4 were comparable to those by ChatGPT-3.5 within the top 10 (43/52, 83% vs 38/52, 73%, respectively; *P*=.34) and top 5 (42/52, 81% vs 34/52, 65%, respectively; *P*=.12) differential diagnosis lists and top diagnosis (31/52, 60% vs 22/52, 42%, respectively; *P*=.12), although the difference was not statistically significant. The rates of correct diagnosis by ChatGPT-4 were also comparable to those by physicians within the top 10 (43/52, 83% vs 39/52, 75%, respectively; *P*=.47) and top 5 (42/52, 81% vs 35/52, 67%, respectively; *P*=.18) differential diagnoses and top diagnosis (31/52, 60% vs 26/52, 50%, respectively; *P*=.43), although the difference was not statistically significant. Multimedia Appendix 2 shows the final diagnosis and the 10 differential diagnosis lists generated by ChatGPT-3.5 and ChatGPT-4 and those created by physicians.

Regarding open access case reports (Table 3), there were no statistically significant differences in the rates of correct diagnoses within the top 10 and top 5 differential diagnosis lists and top diagnoses generated by ChatGPT-3.5 and ChatGPT-4. Regarding the timing of publication, there were no statistically significant differences in ChatGPT-3.5 and ChatGPT-4 diagnoses between those published prior to 2021 and those published in 2022 (Table 3).

Table 1. Representative examples of differential diagnosis lists with the final diagnosis.

Final diagnosis	Differential diagnosis lists		
	ChatGPT-3.5[a]	ChatGPT-4[b]	Physicians
Tarsal tunnel syndrome	• Peripheral neuropathy • Tarsal tunnel syndrome • Lumbar spinal stenosis • Diabetes • Alcohol abuse • Vitamin B12 deficiency • Hypothyroidism • Chronic kidney disease • Multiple sclerosis • Charcot-Marie-Tooth disease	• Peripheral neuropathy • Tarsal tunnel syndrome • Lumbar radiculopathy • Small fiber neuropathy • Chronic inflammatory demyelinating polyneuropathy • Complex regional pain syndrome • Peripheral artery disease • Charcot-Marie-Tooth disease • Polyneuropathy • Leprosy	• Sarcoidosis • Chronic inflammatory demyelinating polyradiculoneuropathy • Diabetic neuropathy • Vitamin B12 deficiency • Alcoholic neuropathy • Paraneoplastic syndrome • Antimyelin oligodendrocyte glycoprotein antibody-associated disease antineutrophil cytoplasmic antibody-related vasculitis • POEMS (polyneuropathy, organomegaly, endocrinopathy, monoclonal plasma cell disorder, skin changes) syndrome • Restless legs syndrome
Nonepisodic angioedema associated with eosinophilia	• Eosinophilic fasciitis • Acute interstitial nephritis • Hypereosinophilic syndrome • Churg-Strauss syndrome • Idiopathic hypereosinophilia syndrome • Lymphoma • Systemic lupus erythematosus • Sarcoidosis • Infectious mononucleosis • Allergic reactions	• Eosinophilic cellulitis (Wells syndrome) • Idiopathic eosinophilic oedema • Hypereosinophilic syndrome • Eosinophilic granulomatosis with polyangiitis (Churg-Strauss syndrome) • Drug-induced eosinophilia • Eosinophilic fasciitis (Shulman syndrome) • Kimura disease • Eosinophilic gastroenteritis • Parasitic infections • Systemic lupus erythematosus	• Hypereosinophilic syndrome • Idiopathic peripheral edema • Idiopathic angioedema • Cholinergic urticaria • Eosinophilic granulomatosis with polyangiitis • Parasitic infection • Chronic eosinophilic leukemia • Drug-induced (including supplements, herbal medicine) hypereosinphilic edema • Tuberculosis • Viral infection
Decompression sickness	• Arterial gas embolism • Decompression sickness • Carbon monoxide poisoning • Nitrogen narcosis • Hypothermia • Acute pancreatitis • Mesenteric ischemia • Gastroenteritis • Gastric ulcer • Acute cholecystitis	• Decompression sickness • Gas embolism • Acute coronary syndrome • Gastroenteritis • Acute pancreatitis • Peptic ulcer disease • Diabetic ketoacidosis • Acute cholecystitis • Mesenteric ischemia • Gastritis	• Air embolization • Decompression sickness • Nonocclusive mesenteric ischemia • Diabetic ketoacidosis • Bacteremia • Bacterial translocation • Acute coronary syndrome • Cholelithiasis • Cholangitis • Cholesterol embolization

[a]ChatGPT-3.5: third-generation ChatGPT.

[b]ChatGPT-4: fourth-generation ChatGPT.

Table 2. Rates of correct diagnoses within the top 10 and top 5 differential diagnosis lists and top diagnosis generated by ChatGPT-3.5 and ChatGPT-4 compared with those created by physicians.

Variable	ChatGPT-4[a] (n=52), n (%)	ChatGPT-3.5[b] (n=52), n (%)	Physicians (n=52), n (%)	*P* value[c]		
				ChatGPT-4 vs physicians	ChatGPT-3.5 vs physicians	ChatGPT-4 vs ChatGPT-3.5
Within the top 10	43 (83)	38 (73)	39 (75)	.47	>.99	.34
Within the top 5	42 (81)	34 (65)	35 (67)	.18	>.99	.12
Top diagnosis	31 (60)	22 (42)	26 (50)	.43	.56	.12

[a]ChatGPT-4: fourth-generation ChatGPT.

[b]ChatGPT-3.5: third-generation ChatGPT.

[c]*P* values from chi-square scores.

Table 3. Rates of correct diagnoses within the top 10 and top 5 differential diagnosis lists and top diagnosis generated by third-generation ChatGPT and fourth-generation ChatGPT between open access and non–open access case reports and between the timing of publications prior to 2021 and published in 2022.

Variable	Fourth-generation ChatGPT						Third-generation ChatGPT					
	Open access (n=39), n (%)	Non–open access (n=13), n (%)	*P* value[a]	Prior to 2021 (n=24), n (%)	In 2022 (n=16), n (%)	*P* value[b]	Open access (n=39), n (%)	Non–open access (n=13), n (%)	*P* value[a]	Prior to 2021 (n=24), n (%)	In 2022 (n=16), n (%)	*P* value[b]
Within the top 10	32 (82)	11 (85)	>.99	20 (83)	13 (81)	>.99	28 (72)	10 (77)	>.99	17 (71)	13 (81)	.71
Within the top 5	31 (80)	11 (85)	>.99	19 (79)	13 (81)	>.99	25 (64)	9 (69)	>.99	17 (71)	11 (69)	>.99
Top diagnosis	22 (56)	9 (69)	.62	17 (71)	9 (56)	.54	14 (36)	8 (62)	.19	11 (46)	8 (50)	>.99

[a]*P* values from chi-square scores comparing open access and non–open access case reports.

[b]*P* values from chi-square scores comparing between case reports prior to 2021 and case reports published in 2022.

Discussion

Principal Results

This study has several main findings. First, our study demonstrates the accuracy of the differential diagnosis lists generated by ChatGPT-3.5 and ChatGPT-4 for complex clinical vignettes from case reports. The rate of correct diagnoses within the top 10 and top 5 differential diagnosis lists generated by ChatGPT-4 was >80%. With a diagnostic accuracy of >80%, ChatGPT-4 can serve as a supplementary tool for physicians, especially when dealing with complex cases. Our results have demonstrated that GPT possesses diagnostic capabilities that can be comparable to those of physicians. This suggests that GPT might serve as a form of collective intelligence, capable of double-checking clinical diagnoses conducted by medical practitioners, at the very least. Second, there were no statistically significant differences in the rates of correct diagnoses by ChatGPT-3.5 and ChatGPT-4 based on the open-access status or the publication date. Both GPT-3.5 and GPT-4 models were constructed using publicly available databases and the knowledge cutoffs set in early 2021 [16,26]. Therefore, we hypothesized that open access case reports could produce better diagnostic results than non–open access ones. Additionally, we postulated that the case reports published in the years prior to 2021 could produce better diagnostic results than the ones published in 2022. The actual results were partly attributed to the limited sample size resulting from the subdivision into exploratory analysis.

Potential Implications for Clinical Practice and Medical Education

The integration of generative AI like ChatGPT into clinical settings could enhance patient care and streamline physician workflows. Given its pretraining accuracy of over 80%, physicians could receive immediate support in challenging cases, thereby minimizing diagnostic errors and enhancing patient outcomes. Furthermore, these AI systems could grant health care professionals more time for the demanding facets of patient care, allowing them to focus on more demanding aspects of patient care and potentially thereby improving health care efficiency. In an educational context, ChatGPT could be

pivotal in shaping future physicians, especially in clinical reasoning and medical knowledge acquisition [27]. Engaging with generative AIs can expose medical learners to an array of diagnoses, preparing them for complex clinical situations.

Limitations

This study has several limitations. First, the study materials were obtained solely from complex case reports published by a single GIM department at a single center. Although these case reports provided insight into challenging diagnostic scenarios, they may not capture the full spectrum of patient presentations, even within the GIM department, as they were not randomly sampled but rather selected for their complexity, unusualness, or the challenges they posed for diagnosis. Therefore, our findings have limited external validity, as they may not be generalizable to other settings. Their performance might differ in simpler or more typical clinical presentations. Second, we acknowledge the possible bias in the differential diagnosis lists. They were created by experienced GIM physicians, implying that the results might not be applicable to lists created by physicians of different specialties or with various levels of training. It would be beneficial if future studies incorporated a wider array of participants. Third, there is a limitation associated with the accessibility and recency of our study. Specifically, 75% (39/52) of the case studies were published as open access, and approximately half of the case studies were published prior to 2021. Although we did not observe statistically significant differences regarding open access and publication timing, there were some possibilities for ChatGPT-3.5, ChatGPT-4, and physicians who created differential diagnosis lists to learn these case materials directly or indirectly. The final limitation pertains to possible time lag when generating differential diagnosis lists between ChatGPT-3.5 and ChatGPT-4. In light of these limitations, future research should assess the diagnostic accuracy of ChatGPT models by using properly tuned case materials that the model has not been trained on.

Comparison With Prior Work

Our previous study [22] showed that the diagnostic accuracy of ChatGPT-3.5 was lower than that of physicians (25/30, 83% vs 59/60, 98%, respectively). In contrast, the findings of this study revealed that the rates of correct diagnoses within the top 10 (43/52, 83% vs 39/52, 75%, respectively) and top 5 (42/52, 81% vs 35/52, 67%, respectively) differential diagnosis lists, as well as the top diagnosis (31/52, 60% vs 26/52, 50%, respectively) generated by ChatGPT-4 were comparable to those by physicians. These results suggest the evolving performance of AI chatbots across different ChatGPT versions. Compared with those in the prior study [22], the rates of correct diagnoses within the top 10 (38/52, 73% vs 28/30, 93%, respectively) and top 5 (34/52, 65% vs 25/30, 83%, respectively) differential diagnosis lists and top diagnosis (22/52, 42% vs 16/30, 53%, respectively) generated by ChatGPT-3 (or 3.5) were lower in this study. This discrepancy was largely attributed to this study's emphasis on complex clinical case vignettes sourced from case reports within the GIM department, while the prior research focused on more common clinical presentations. Moreover, ChatGPT-4 provided better results in its differential diagnosis lists (43/52, 83% vs 45/70, 64%, respectively) and as its top diagnosis (31/52, 60% vs 27/70, 39%, respectively) compared with those reported in another study for New England Journal of Medicine clinicopathologic conferences [21]. These variations can be partly ascribed to differences in the study designs, including case vignettes and systems.

Compared with a previous review on symptom checkers [5], the rate of correct diagnoses within the top 10 differential diagnoses generated by ChatGPT-4 was higher (43/52, 83% vs 60.9%-76.9%, respectively) in this study. Compared with a previous review on the differential diagnosis generator [6], the rate of correct diagnoses within the top 10 differential diagnoses generated by ChatGPT-4 was higher (43/52, 83% vs 63%-77%, respectively) in this study. This discrepancy is partly due to differences in study designs, case materials, and algorithms. In the future, direct comparisons between ChatGPT and other CDS systems are required.

Conclusions

This study demonstrates the potential diagnostic accuracy of the differential diagnosis lists generated by ChatGPT-3.5 and ChatGPT-4 by using complex clinical vignettes from case reports published by the GIM department. Notably, the rate of correct diagnoses within the top 10 and top 5 differential diagnosis lists generated by ChatGPT-4 exceeds 80%. Although these results stem from a limited data set of case reports from a single department, they indicate the potential utility of ChatGPT-4 as a supplementary tool for physicians, particularly for those affiliated with the GIM department. Future research should assess the diagnostic accuracy of ChatGPT models by using properly tuned case materials that the model has not been trained on. Additionally, future investigations should evaluate the literacy level of AIs and their alignment with relevant medical text. Such efforts will ensure a comprehensive insight into the AI's possible roles in enhancing clinical decision-making processes. Moreover, as AI systems become more prevalent, their influence is expected to ripple across various facets of health care. Generative AIs have the potential to reshape patient-physician dynamics, fostering more informed interactions. They can also play a pivotal role in democratizing medical knowledge. This could lead to heightened health care accessibility, allowing even those in remote or underserved regions to glean expert medical advice. Given these profound implications, it becomes imperative to investigate the ramifications of AI integration into health care.

Authors' Contributions

TH, RK, YH, KM, KT, YK, T Suzuki, and T Shimizu contributed to the study concept and design. TH performed the statistical analyses. TH contributed to the drafting of the manuscript. RK, YH, KM, KT, YK, T Suzuki, and T Shimizu contributed to the critical revision of the manuscript for relevant intellectual content. All the authors have read and approved the final version of the manuscript. We would like to specially thank Dr Kenjiro Kakimoto, Department of Psychiatry, Nihon University School of

Medicine, for helping us with the analysis. This study was conducted using resources from the Department of Diagnostics and Generalist Medicine at Dokkyo Medical University.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Case reports included in this study.
[PDF File (Adobe PDF File), 52 KB - medinform_v11i1e48808_app1.pdf]

Multimedia Appendix 2
Final diagnosis and the differential diagnosis lists generated by ChatGPT and those created by physicians.
[PDF File (Adobe PDF File), 203 KB - medinform_v11i1e48808_app2.pdf]

References

1. Holmboe E, Durning S. Assessing clinical reasoning: moving from in vitro to in vivo. Diagnosis (Berl) 2014 Jan 01;1(1):111-117 [FREE Full text] [doi: 10.1515/dx-2013-0029] [Medline: 29539977]
2. Harada Y, Otaka Y, Katsukura S, Shimizu T. Effect of contextual factors on the prevalence of diagnostic errors among patients managed by physicians of the same specialty: a single-centre retrospective observational study. BMJ Qual Saf 2023 Jan 23:bmjqs-2022-015436. [doi: 10.1136/bmjqs-2022-015436]
3. Skinner T, Scott I, Martin J. Diagnostic errors in older patients: a systematic review of incidence and potential causes in seven prevalent diseases. IJGM 2016 May:137-146. [doi: 10.2147/ijgm.s96741] [Medline: 27284262]
4. Committee on Diagnostic Error in Health Care, Board on Health Care Services, Balogh EP, Miller BT. Technology and tools in the diagnostic process. In: Improving Diagnosis in Health Care. Washington DC: National Academies Press (US); Dec 29, 2015.
5. Schmieding ML, Kopka M, Schmidt K, Schulz-Niethammer S, Balzer F, Feufel MA. Triage accuracy of symptom checker apps: 5-year follow-up evaluation. J Med Internet Res 2022 May 10;24(5):e31810 [FREE Full text] [doi: 10.2196/31810] [Medline: 35536633]
6. Riches N, Panagioti M, Alam R, Cheraghi-Sohi S, Campbell S, Esmail A, et al. The effectiveness of electronic differential diagnoses (ddx) generators: a systematic review and meta-analysis. PLoS One 2016;11(3):e0148991 [FREE Full text] [doi: 10.1371/journal.pone.0148991] [Medline: 26954234]
7. Greenes R. Chapter 2 - A brief history of clinical decision support: technical, social, cultural, economic, governmental perspectives. In: Clinical Decision Support (Second Edition). London, UK: Academic Press; Mar 28, 2014:49-109.
8. Kawamura R, Harada Y, Sugimoto S, Nagase Y, Katsukura S, Shimizu T. Incidence of diagnostic errors among unexpectedly hospitalized patients using an automated medical history-taking system with a differential diagnosis generator: retrospective observational study. JMIR Med Inform 2022 Jan 27;10(1):e35225 [FREE Full text] [doi: 10.2196/35225] [Medline: 35084347]
9. Meunier P, Raynaud C, Guimaraes E, Gueyffier F, Letrilliart L. Barriers and facilitators to the use of clinical decision support systems in primary care: a mixed-methods systematic review. Ann Fam Med 2023;21(1):57-69 [FREE Full text] [doi: 10.1370/afm.2908] [Medline: 36690490]
10. Wani SUD, Khan NA, Thakur G, Gautam SP, Ali M, Alam P, et al. Utilization of artificial intelligence in disease prevention: diagnosis, treatment, and implications for the healthcare workforce. Healthcare (Basel) 2022 Mar 24;10(4):608 [FREE Full text] [doi: 10.3390/healthcare10040608] [Medline: 35455786]
11. Haug CJ, Drazen JM. Artificial intelligence and machine learning in clinical medicine, 2023. N Engl J Med 2023 Mar 30;388(13):1201-1208. [doi: 10.1056/nejmra2302038]
12. No authors listed. Will ChatGPT transform healthcare? Nat Med 2023 Mar;29(3):505-506. [doi: 10.1038/s41591-023-02289-5] [Medline: 36918736]
13. Biswas S. ChatGPT and the future of medical writing. Radiology 2023 Apr;307(2):e223312. [doi: 10.1148/radiol.223312] [Medline: 36728748]
14. Curtis N, ChatGPT. To ChatGPT or not to ChatGPT? The impact of artificial intelligence on academic publishing. Pediatr Infect Dis J 2023 Apr 01;42(4):275. [doi: 10.1097/INF.0000000000003852] [Medline: 36757192]
15. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb;2(2):e0000198 [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]
16. OpenAI. GPT-4 technical report. arXiv Preprint posted online on March 15, 2023 [FREE Full text] [doi: 10.48550/arXiv.2303.08774]
17. Vaishya R, Misra A, Vaish A. ChatGPT: Is this version good for healthcare and research? Diabetes Metab Syndr 2023 Apr;17(4):102744. [doi: 10.1016/j.dsx.2023.102744] [Medline: 36989584]

18. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med 2023 Mar 30;388(13):1233-1239. [doi: 10.1056/nejmsr2214184]
19. Zheng H, Zhan H. ChatGPT in scientific writing: A cautionary tale. Am J Med 2023 Aug;136(8):725-726.e6. [doi: 10.1016/j.amjmed.2023.02.011] [Medline: 36906169]
20. Chen JH, Dhaliwal G, Yang D. Decoding artificial intelligence to achieve diagnostic excellence: learning from experts, examples, and experience. JAMA 2022 Aug 23;328(8):709-710. [doi: 10.1001/jama.2022.13735] [Medline: 35913752]
21. Kanjee Z, Crowe B, Rodman A. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. JAMA 2023 Jul 03;330(1):78-80. [doi: 10.1001/jama.2023.8288] [Medline: 37318797]
22. Hirosawa T, Harada Y, Yokose M, Sakamoto T, Kawamura R, Shimizu T. Diagnostic accuracy of differential-diagnosis lists generated by generative pretrained transformer 3 chatbot for clinical vignettes with common chief complaints: a pilot study. Int J Environ Res Public Health 2023 Feb 15;20(4):3378 [FREE Full text] [doi: 10.3390/ijerph20043378] [Medline: 36834073]
23. Jinno A, Hirosawa T, Shimizu T. Hepatic portal venous gas after diving. BMJ Case Reports 2018 Jan 12:bcr-2017-223844. [doi: 10.1136/bcr-2017-223844]
24. Krupat E, Wormwood J, Schwartzstein RM, Richards JB. Avoiding premature closure and reaching diagnostic accuracy: some key predictive factors. Med Educ 2017 Nov;51(11):1127-1137. [doi: 10.1111/medu.13382] [Medline: 28857266]
25. Armstrong RA. When to use the Bonferroni correction. Ophthalmic Physiol Opt 2014 Sep;34(5):502-508. [doi: 10.1111/opo.12131] [Medline: 24697967]
26. Zong M, Krishnamachari B. A survey on GPT-3. arXiv Preprint posted on December 1, 2022. [doi: 10.48550/arXiv.2212.00857]
27. Hirosawa T, Shimizu T. Enhancing clinical reasoning with chat generative pre-trained transformer: A practical guide. Diagnosis 2023;11(1):A. [doi: 10.1515/dx-2023-0116]

Abbreviations

AI: artificial intelligence
CDS: clinical decision support
GIM: general internal medicine
GPT: generative pretrained transformer
GPT-3.5: third-generation generative pretrained transformer
GPT-4: fourth-generation generative pretrained transformer

Edited by A Castonguay; submitted 09.05.23; peer-reviewed by D Chrimes, M Kopka; comments to author 17.07.23; revised version received 20.07.23; accepted 13.09.23; published 09.10.23.

Please cite as:
Hirosawa T, Kawamura R, Harada Y, Mizuta K, Tokumasu K, Kaji Y, Suzuki T, Shimizu T
ChatGPT-Generated Differential Diagnosis Lists for Complex Case–Derived Clinical Vignettes: Diagnostic Accuracy Evaluation
JMIR Med Inform 2023;11:e48808
URL: https://medinform.jmir.org/2023/1/e48808
doi:10.2196/48808
PMID:37812468

Establishment of a Public Mental Health Database for Research Purposes in the Ferrara Province: Development and Preliminary Evaluation Study

Maria Ferrara[1,2,3], MD, PhD; Elisabetta Gentili[4], MA; Martino Belvederi Murri[1,2], MD; Riccardo Zese[5], PhD; Marco Alberti[6], PhD; Giorgia Franchini[7], PhD; Ilaria Domenicano[1], PhD; Federica Folesani[1,2], MD; Cristina Sorio[2], MPH; Lorenzo Benini[2], BA; Paola Carozza[2], MD; Julian Little[8], PhD; Luigi Grassi[1,2], MD

1

2

3

4

5

6

7

8

Corresponding Author:
Maria Ferrara, MD, PhD

Abstract

Background: The immediate use of data exported from electronic health records (EHRs) for research is often limited by the necessity to transform data elements into an actual data set.

Objective: This paper describes the methodology for establishing a data set that originated from an EHR registry that included clinical, health service, and sociodemographic information.

Methods: The Extract, Transform, Load process was applied to raw data collected at the Integrated Department of Mental Health and Pathological Addictions in Ferrara, Italy, from 1925 to February 18, 2021, to build the new, anonymized Ferrara-Psychiatry (FEPSY) database. Information collected before the first EHR was implemented (ie, in 1991) was excluded. An unsupervised cluster analysis was performed to identify patient subgroups to support the proof of concept.

Results: The FEPSY database included 3,861,432 records on 46,222 patients. Since 1991, each year, a median of 1404 (IQR 1117.5-1757.7) patients had newly accessed care, and a median of 7300 (IQR 6109.5-9397.5) patients were actively receiving care. Among 38,022 patients with a mental disorder, 2 clusters were identified; the first predominantly included male patients who were aged 25 to 34 years at first presentation and were living with their parents, and the second predominantly included female patients who were aged 35 to 44 years and were living with their own families.

Conclusions: The process for building the FEPSY database proved to be robust and replicable with similar health care data, even when they were not originally conceived for research purposes. The FEPSY database will enable future in-depth analyses regarding the epidemiology and social determinants of mental disorders, access to mental health care, and resource utilization.

(JMIR Med Inform 2023;11:e45523) doi:10.2196/45523

KEYWORDS

mental health; psychosis; epidemiology; electronic health registry; health care; machine learning; medical health records; electronic health records; clinical database; support; mental disorder; social determinants; mental health care; resource utilization

Introduction

Electronic health records (EHRs) assemble and enable access to large volumes of clinical and sociodemographic data that are routinely collected by local health authorities. EHRs offer a unique opportunity to conduct research on various topics, including, among others, the patterns of health care resource use and factors that influence the course and outcomes of mental disorders in large, representative samples [1,2]. EHRs can be linked to data related to census and geolocalization information [3]; such investigations span the epidemiology of mental disorders, hospitalization rates, morbidity, and mortality.

The breadth and nature of information represented in the sample of EHRs in the mental health sector make such information

particularly suitable for using artificial intelligence (AI) and machine learning techniques, in addition to traditional methods (eg, linear regression models), in order to increase the potential for research on social and clinical factors [4].

Applications that use AI take advantage of AI's ability to process large amounts of data in order to extract information or identify underlying patterns of relationships that conventional methods may overlook [5]. AI may be particularly suitable for the investigation of large amounts of clinical data, thanks to (1) the flexibility and scalability of AI techniques, which are higher than those of traditional methods, and (2) the ability of AI to consider all of the available predictors (ie, not only a subset), which makes AI and, in particular, machine learning suitable for performing tasks such as classification, prediction, and resource optimization [5,6].

Indeed, in recent years, the use of AI techniques in mental health care research have rapidly increased, including its use to identify a disease at its earliest stages, predict illness onset in vulnerable individuals, study illness progression, optimize treatment, and discover novel therapeutic agents [7,8].

As of yet, there are few examples (mainly from the United States) of how data collected from EHRs can be successfully adapted for analysis with AI. For example, Hughes et al [9] analyzed clinical variables in the EHRs of 81,630 adults from 2 academic medical centers in Boston, Massachusetts (spanning 10 years) and identified predictors of treatment response for major depressive disorders.

Xu et al [10] compiled a data set of 11,275 patients from 5 large medical centers across New York City by using EHR data collected between 2008 and 2017; they used machine learning methods to identify markers of depression phenotypes to inform clinical decisions about patients' care. Pradier et al [11] analyzed a data set of 67,807 individuals to predict the risk of receiving a misdiagnosis of bipolar disorder among individuals with depression, using only information retrieved from EHRs. Perlis et al [12] applied natural language processing to classify the mood states of 127,504 patients, using data from an EHR.

In order to fully exploit the potential of EHRs for mental health research however, important issues need to be considered. One preliminary, controversial issue is whether the use of EHR data should be restricted to the purpose for which they were collected [13]. Indeed, privacy constraints, data security, and overall ethics regulation must be taken into account when considering whether to use EHRs for research purposes [6]. Nonetheless, nowadays, medical data that were originally collected for purposes other than research are being used to study health phenomena in many different fields, including mental health, substance use, noncommunicable diseases (eg, cancer), and health behaviors (eg, cancer screening) [14]. A further challenge is that data may not be homogeneous or may not be collected systematically, and most data are not derived from structured scales or questionnaires. The adaptation of the EHR represents the first necessary step to planning research projects that include models for predicting health resource utilization, identifying predictors of diagnostic accuracy, and differentiating between remission and chronicity, as done in other fields such as oncology [15].

Given this premise, the aim of this paper is to describe (1) the challenges and pitfalls that were encountered in the process of adapting EHR data derived from the public mental health agency in Ferrara, Italy, for research purposes and (2) the development of a data set that is suitable for analysis via AI and traditional techniques. In order to test the feasibility of using these data in analyses and the robustness of analyses based on such data, a clustering analysis was also performed, and preliminary results are presented herein.

Methods

Ethics Approval

Ethical approval was obtained by the Area Vasta Emilia Centro Ethical Committee on December 12, 2019 (protocol number: 197/2018/Oss/AUSLF). This study conforms to the principles expressed in the Declaration of Helsinki.

Setting

In Italy, mental health care is provided by departments of mental health [16-19]. The levels of care within each department of mental health include community-based mental health centers, hospital psychiatric inpatient units, and rehabilitation or residential facilities. Each community-based mental health center serves as a hub of psychiatric care for geographically defined catchment areas with 50,000 to 150,000 inhabitants [20] In Ferrara, Northern Italy, the Integrated Department of Mental Health and Pathological Addiction covers an area of 2630 km^2, with a catchment of 342,061 inhabitants as of 2020 [21].

Data Collection and EHRs

Data were collected in 2 periods that were distinct in terms of the methodology used, the psychiatric services delivered, and the level of digitalization. Data related to the first period, which began in 1925 and ended in 1990, were gathered mostly in a psychiatric asylum, during a time when digital health was not fully developed or adopted.

In 1991, the first structured EHR (ie, SIPER [Sistema Informativo Psichiatrico dell'Emilia-Romagna]) was introduced and implemented locally by the Local Health Trust of Ferrara for Mental Health in Adults. Different software programs were adopted during the years following the implementation of SIPER, and each new software program replaced the previous one by importing already existing data and adding new features (and thus information), as detailed in Textbox 1.

Textbox 1. Electronic health records implemented by the Local Health Trust of Ferrara for Mental Health in Adults in chronological order.

1. SIPER (Sistema Informativo Psichiatrico dell'Emilia-Romagna; 1991-1994) included individual demographic data, medical records, diagnoses, and health services.
2. CINECA (1994-1998) added the feature labeled as "project," which was defined as the comprehensive set of treatments and activities offered to the patient.
3. GESAP (Gestione attività Psichiatrica; 1998-2004) added information about outpatient treatment; hospitalizations in inpatient units, long-term residences, and semiresidences; and outpatient services.
4. IPPOCRATE (GPI SRL; 2004-2008).
5. EFESO (Newteam SRL; 2008-2021) added the text field labeled "evaluation and treatment area" in the medical record, structured diagnostic evaluation, pharmacological treatment prescription and administration, clinical notes, attached documents, and a feature to identify structured clinical protocols (Percorso Diagnostico Terapeutico Assistenziale; diagnostic and therapeutic care pathway).
6. CURE (Cartella Unificata Regionale Elettronica; Engineering SpA; 2021 to present) added the registration of vital signs and laboratory tests, as well as legal and administrative documentation.

Data Preparation

The first goal was the creation of a new, fully deidentified database with data available in EFESO (Newteam SRL)—a necessary step for complying with privacy constraints.

In order to remove all protected health information (PHI), source data needed to be modified. This could not be done by directly editing data in EFESO, since the source could not be altered directly. Thus, the new research database—the Ferrara-Psychiatry (FEPSY) database—was built via the Extract, Transform, Load process, which is a 3-phase process [22] in which data are first extracted from 1 source or multiple and possibly different sources (eg, databases, flat or formatted files, and web pages). Afterward, the extracted data are stored in a staging area, where they undergo transformation, such as filtering, cleaning, summarization, and normalization. Finally, the transformed data are loaded into the destination storage. For example, one type of transformation was record exclusion. We excluded records of patients that could not be unequivocally identified by the tax code—a unique 16-digit alphanumeric code that identifies a person in Italian public administration forms.

While assessing the suitability of the FEPSY database for research purposes, we noted that historical information dating back up to 1925 had also been maintained in EFESO. We understood that such data were manually imported into the electronic databases that preceded SIPER (year 1991); however, because we could not confirm the procedures, scope, and quality of this historical data import, we decided to document the existence of these data but exclude them from analysis.

More details about the FEPSY database can be found in Multimedia Appendix 1. In Table S1 in Multimedia Appendix 1, for each table in the FEPSY database, the total number of rows (number and percentage of records retained in the FEPSY database) is reported, alongside the number of records in the corresponding original EFESO table from which data were extracted (number of records in the corresponding EFESO table). As detailed in Table S1 in Multimedia Appendix 1, of the 4,264,954 records, 3,861,432 (90.54%) were kept. These records included detailed information about the patient, their illness, and the treatments provided.

In Table S2 in Multimedia Appendix 1, 2 types of anomalies for each table in the FEPSY database are described; one is date inconsistency (eg, when the closing date precedes the opening date of the medical chart), and the other is a date anomaly that was generated by the automatized mechanism that was introduced by EFESO when migrating data from IPPOCRATE (GPI SRL; August 26, 2008).

Clustering

Once the anonymized database was built, a clustering analysis was performed to investigate the data set quality. A clustering algorithm is an unsupervised machine learning technique that is used to group objects, so that objects of the same group (or cluster) are very similar to one another and objects of different groups are very dissimilar. To decide the degree of similarity (or dissimilarity) between 2 objects, various distance measures can be used, such as the Euclidean distance between (normalized) numerical representations of the objects. We tested the hypothesis that the patients modeled in the data set could be divided into homogeneous clusters. The k-means algorithm computes a numerical distance between objects to determine to which cluster they belong. However, in our case, data were categorical. In 1995, Ralambondrainy [23] introduced an approach that enables the use of the k-means algorithm with categorical data. In this approach, nominal attributes are converted into binary attributes—one for each value that the attribute can take—so that they can be considered as numerical attributes by the algorithm.

We performed the clustering analysis with the WEKA (Waikato Environment for Knowledge Analysis; University of Waikato) data mining tool, which provides an implementation of the k-means algorithm (ie, SimpleKMeans) and can handle categorical data [24]. SimpleKMeans can also handle missing values by replacing them with the mean or mode.

For this preliminary clustering analysis, we included only the patients who had at least one recorded diagnosis of a mental disorder (ie, *International Classification of Diseases, Ninth Revision* [*ICD-9*] codes 290-319) [25].

Patients were excluded if they had nonpsychiatric diagnoses (*ICD-9* codes V01-V91; 2707/46,222, 5.86%) or had never received an *ICD-9* diagnosis (5493/46,222, 11.88%). The

resulting subset for the clustering analysis included 38,022 patients.

We considered sociodemographic variables, such as biological sex, age at first visit, nationality, marital status, living situation, education, occupational status, birthplace (district), and the catchment area (district) providing care (determined by domicile postal code or by residence postal code when the domicile was missing).

Results

Sociodemographic Characteristics

The sample included 46,222 individuals, whose sociodemographic characteristics are detailed in Table 1.

Table . Sociodemographic characteristics of all of the individuals who accessed mental health services in the Ferrara province (1991-2021) and were included in the FEPSY[a] database.

Characteristic		Female patients (n=28,109)	Male patients (n=18,113)	All patients (N=46,222)
Age at first visit (years), mean (SD)		50.46 (18.82)	48.72 (19.02)	49.78 (18.91)
	<18, n (%)	249 (0.89)	164 (0.91)	413 (0.89)
	18-24, n (%)	2173 (7.73)	1842 (10.17)	4015 (8.69)
	25-34, n (%)	4221 (15.02)	2913 (16.08)	7134 (15.43)
	35-44, n (%)	5141 (18.29)	3308 (18.26)	8449 (18.28)
	45-54, n (%)	4761 (16.94)	3119 (17.22)	7880 (17.05)
	55-64, n (%)	4142 (14.74)	2423 (13.38)	6565 (14.20)
	65-74, n (%)	3780 (13.45)	2180 (12.04)	5960 (12.89)
	≥75, n (%)	3634 (12.93)	2160 (11.93)	5794 (12.54)
	Missing data, n (%)	8 (0.03)	4 (0.02)	12 (0.03)
Nationality, n (%)				
	Italian	26,486 (94.23)	17,167 (94.78)	43,653 (94.44)
	Foreign	1580 (5.62)	920 (5.08)	2500 (5.41)
	Missing data	43 (0.15)	26 (0.14)	69 (0.15)
Birthplace (district), n (%)				
	Outside Ferrara province	7525 (26.77)	4825 (26.64)	12,350 (26.72)
	Ferrara	7299 (25.97)	5034 (27.79)	12,333 (26.68)
	Codigoro	3414 (12.15)	2167 (11.96)	5581 (12.07)
	Portomaggiore	2803 (9.97)	1795 (9.91)	4598 (9.95)
	Copparo	2439 (8.68)	1466 (8.09)	3905 (8.45)
	Cento	2262 (8.05)	1463 (8.08)	3725 (8.06)
	Outside Italy	2253 (8.02)	1258 (6.95)	3511 (7.60)
	Missing data	114 (0.41)	105 (0.58)	219 (0.47)
Marital status, n (%)				
	Married or partnered	10,748 (38.24)	6001 (33.13)	16,749 (36.24)
	Single	5551 (19.75)	5721 (31.59)	11,272 (24.39)
	Separated, divorced, or widowed	5587 (19.88)	1744 (9.63)	7331 (15.86)
	Missing data	6223 (22.14)	4647 (25.66)	10,870 (23.52)
Living situation, n (%)				
	Living with acquired family (partner and children)	12,147 (43.21)	6148 (33.94)	18,295 (39.58)
	Living with parents	3079 (10.95)	3338 (18.43)	6417 (13.88)
	Alone	3194 (11.36)	1768 (9.76)	4962 (10.74)
	Living with other family members	1649 (5.87)	781 (4.31)	2430 (5.26)
	Living with others (eg, roommates)	612 (2.18)	403 (2.22)	1015 (2.20)
	Community housing facilities	184 (0.65)	267 (1.47)	451 (0.98)
	Other	193 (0.69)	220 (1.21)	413 (0.89)
	Safe house	181 (0.64)	199 (1.10)	380 (0.82)
	Retirement home	226 (0.8)	146 (0.81)	372 (0.80)

Characteristic		Female patients (n=28,109)	Male patients (n=18,113)	All patients (N=46,222)
	Prison	1 (3.56×10^{-5})	16 (0.09)	17 (0.04)
	Missing data	6643 (23.63)	4827 (26.65)	11,470 (24.82)
Education, n (%)				
	Illiterate	2941 (10.46)	1540 (8.50)	4481 (9.69)
	Literate (without formal degree)	3044 (10.83)	2312 (12.76)	5356 (11.59)
	Primary school	3674 (13.07)	2167 (11.96)	5841 (12.64)
	Middle school	3247 (11.55)	2644 (14.60)	5891 (12.75)
	High school	3946 (14.04)	2378 (13.13)	6324 (13.68)
	College or university	1294 (4.60)	584 (3.22)	1878 (4.06)
	Missing data	9963 (35.44)	6488 (35.82)	16,451 (35.59)
Occupational status, n (%)				
	Employed	3873 (13.78)	2735 (15.10)	6608 (14.30)
	Retired	2949 (10.49)	1833 (10.12)	4782 (10.35)
	Unemployed	1531 (5.45)	1291 (7.13)	2822 (6.11)
	Disability	612 (2.18)	670 (3.70)	1282 (2.77)
	Other	760 (2.70)	491 (2.71)	1251 (2.71)
	Homemaker	944 (3.36)	1 (0.01)	945 (2.04)
	Student	512 (1.82)	334 (1.84)	846 (1.83)
	Unknown	16,928 (60.22)	10,758 (59.39)	27,686 (59.90)
Catchment area (district), n (%)				
	Ferrara	10,964 (39.01)	6784 (37.45)	17,748 (38.40)
	Codigoro	4186 (14.89)	2931 (16.18)	7117 (15.40)
	Portomaggiore	3425 (12.18)	2193 (12.11)	5618 (12.15)
	Cento	3380 (12.02)	2097 (11.58)	5477 (11.85)
	Copparo	2754 (9.80)	1671 (9.23)	4425 (9.57)
	Unknown	3400 (12.10)	2437 (13.45)	5837 (12.63)

[a]FEPSY: Ferrara-Psychiatry.

Extract, Transform, Load Process

Built-In Tables

Built-in tables from the EFESO relational database, which are detailed in Textbox 2, were included in the FEPSY database.

Textbox 2. Built-in tables included in the FEPSY (Ferrara-Psychiatry) database.

1. Table *Patients* included individual personal data, such as name, place and date of birth, biological sex (male or female), home address, living condition, education, marital status, occupation, and other sociodemographic characteristics.
2. Table *Medical Records* contained 1 or more medical records for each patient, with information such as the date of admission, date and type of discharge, primary diagnosis of a mental disorder, and facility providing care.
3. Table *Diagnoses* included 1 or more diagnoses that were assigned to each individual. Diagnoses were classified according to the *International Classification of Diseases, Ninth Revision* (*ICD-9*) categorical system; therefore, every diagnosis included the associated *ICD-9* code, description, group, and chapter [24]. Diagnoses recorded before the introduction of the *ICD-9* were recorded in SIPER (Sistema Informativo Psichiatrico dell'Emilia-Romagna), using standardized conversion criteria [24].
4. Table *Products* referred to the different types of medical services, such as consultations or hospitalizations. A product had a start date and end date, and it may have contained 1 or more medical services.
5. Table *Medical Services* stored every service that each individual had received or undergone, such as consultations, first visits, the administration of pharmacological treatment, social skill–oriented activities, structured diagnostic assessments, and mandatory medical treatments, as well as the facility providing care.
6. Tables *Medication Prescription* and *Medication Administration* referred to the prescription and administration of pharmacological treatment, type of medication and dosage, start and stop dates, and responsible facility.
7. Table *Psychometric Tests* included every test administered to each patient and the test types, dates, questions, and scores.
8. Table *Projects* listed the treatment plans for each patient. There were individual and group projects, and within a project, there could have been 1 or more products and medical services.
9. Table *Facilities* contained all of the facilities of the Health Trust of Ferrara for Mental Health in Adults, such as hospitals, day care centers, and clinics, along with their types and locations.

Extract

Data were extracted from EFESO by using an automated procedure that executed an SQL select query. This query selected all relevant fields of a table and other useful information from linked support tables, such as the descriptions of the codes. The result of the query was stored in a Pandas DataFrame (The Pandas Development Team) [26,27], which can be easily manipulated in the next phase.

Transform

Record Exclusion

Data imported before 1991 were excluded (as detailed in the *Data Preparation* section). Some fields and records were removed [22] to ensure data consistency, because there were duplicate or erroneous records (Table S1 in Multimedia Appendix 1). These were (1) fields containing unreliable information, (2) fields that were present but not in use (their values were always null), and (3) all records marked as "deleted" (ie, wrong records that were not to be used) and all records in other tables referencing the "deleted" ones. We decided to remove 36 individuals that had unique fiscal codes but duplicate patient IDs—corresponding to 0.15% (72/48,001) of the records in the total data set—since it was not possible to determine which of the two entries was the correct one. When a patient was first included in the database, a unique identifier—the patient ID—was assigned. The combination of the tax code and the patient ID allowed for the unique identification of a patient in the database. We also excluded patients for whom a record was opened earlier than the birth date (16/48,001, 0.03%), patients with no medical records (603/48,001, 1.26%), and patients marked as "deleted" (77/48,001, 0.16%). Overall, 1.6% (768/48,001) of the total records, which related to 732 patients, were removed from the source table.

Anonymization

Anonymization was necessary in order to use the extracted data for research projects and was performed on tables *Patients* and *Medical Records*. First, the extracted records were shuffled. Afterward, the original patient and medical record IDs were replaced with a universally unique identifier (UUID), which is a 128-bit string that is usually represented as a sequence of 32 hexadecimal digits [28]. These new random, unique identifiers were generated with the *uuid4* function of the Python *uuid* package (Python Software Foundation) [29] and used as the new primary key. In order to maintain the referential integrity (ie, the primary key of one table is a foreign key in another table, meaning that they are related), the old IDs were replaced with the new ones within every table in which they appeared. Furthermore, all PHI were excluded from table *Patients*; these data included first names, last names, days and months of birth, tax codes, home addresses, phone numbers, and note fields that could potentially include personal data (eg, relatives' names). For the same reason, text note fields were also excluded from other tables, when present.

Field Transformation

Transformation was necessary for date fields. EFESO stored dates in "datetime" format, that is, "dd/mm/yyyy hh:mm." However, previous EHRs stored only the date, without the hour information. Furthermore, even when specified, the hour information is not always reliable. For this reason, the date fields were split into 2 fields—one for the date and one for the time.

Missing Values

Missing values were assessed to avoid the introduction of bias. In specific analyses, the level and pattern of missingness will be assessed for each variable included and dealt with accordingly.

Load

Data extracted from EFESO were loaded in the FEPSY database—the newly created MySQL (Oracle Corporation) relational database—by using the same automated procedure that was used to extract them. For each built-in table, an insert query, which took the values from the same DataFrame of the select query, was executed.

Analysis of the Extracted Data

The data included in the final composite FEPSY data set were those collected from 1991 to February 2021. Since 1991, each year, a median of 1404 (IQR 1117.5-1757.7) individuals had newly accessed care, and a median of 7300 (IQR 6109.5-9397.5) individuals were actively receiving care, as represented in Figure 1. Figure 2 shows the number of patients treated per year in total and by sex. The sudden decrease observed in 2009 was due to an automated closing procedure that was introduced in 2008 and retained from then on. When migrating from IPPOCRATE to EFESO, all medical records, products, and diagnoses that had not been updated in 365 days were assumed to be closed or terminated, and the missing closing date was replaced with the date of the migration to EFESO, that is, August 26, 2008.

Figure 1. New admissions per year in total and by sex (upper panel). Timeline of the electronic health records adopted by the health care agency in the Ferrara province (lower panel). CURE: Cartella Unificata Regionale Elettronica; GESAP: Gestione attività Psichiatrica; SIPER: Sistema Informativo Psichiatrico dell'Emilia-Romagna.

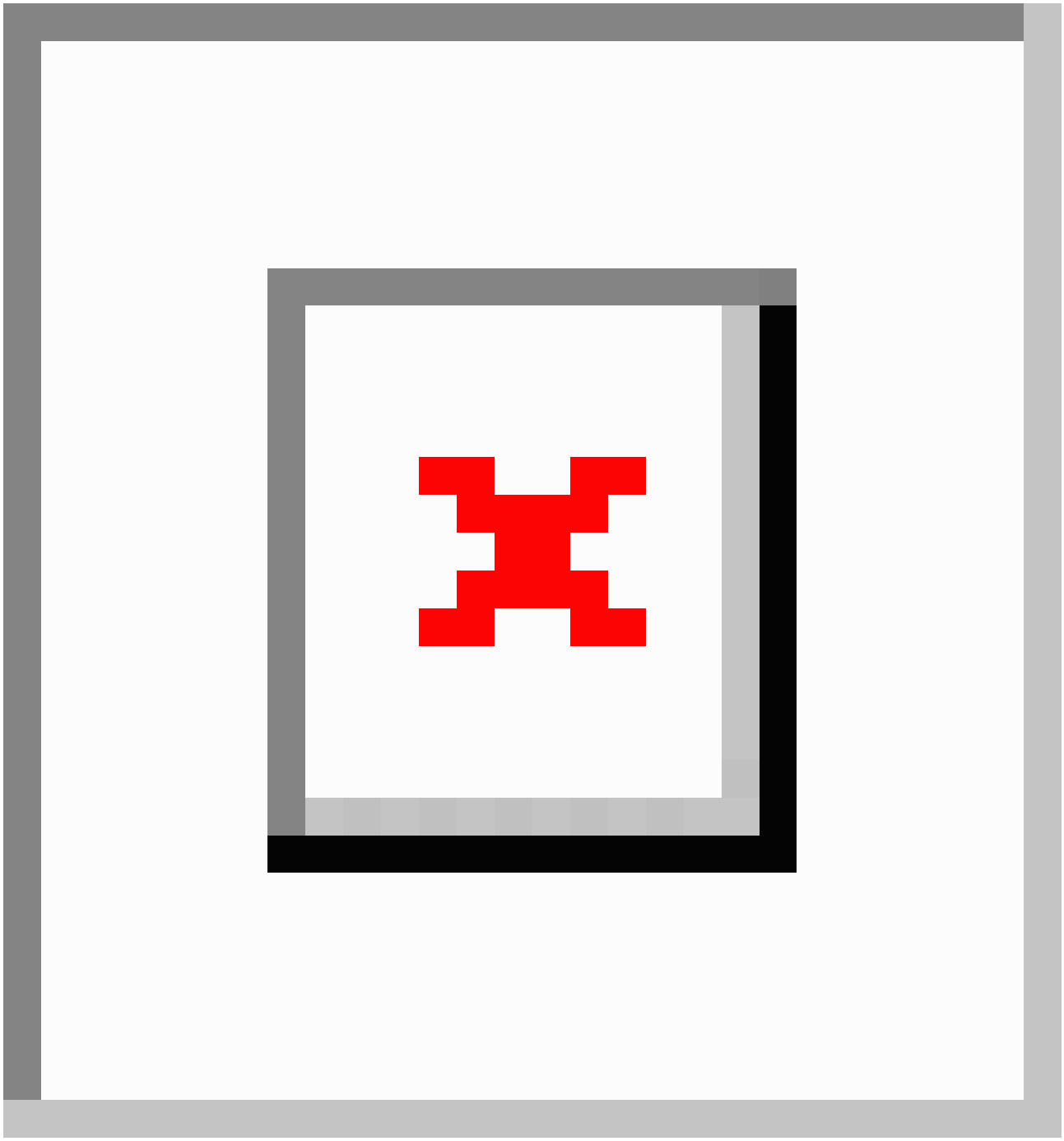
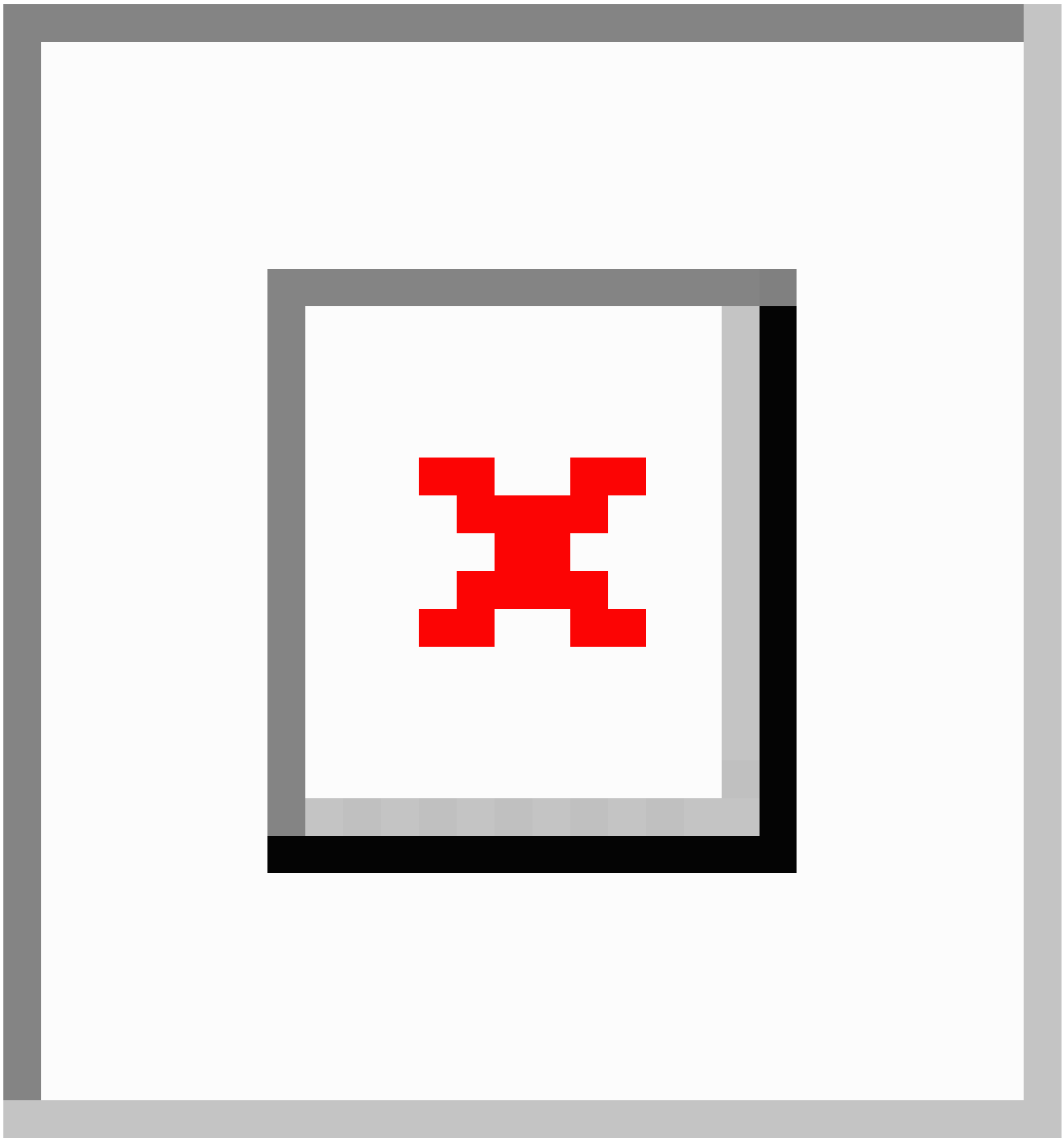


Figure 2. Patients receiving care from the mental health services in Ferrara over time (years 1991-2021), in total (continuous line) and by sex (dotted and dashed lines). CURE: Cartella Unificata Regionale Elettronica; GESAP: Gestione attività Psichiatrica. SIPER: Sistema Informativo Psichiatrico dell'Emilia-Romagna.



As described in Table 2, the most frequent diagnoses at first admission were depression and anxiety disorder. During the 30-year time span, more than half (32,230/46,222, 69.73%) of the patients had only 1 chart open, and only 5184 patients had at least one psychiatric hospitalization.

Table . Main clinical characteristics of the sample (N=46,222; years 1991-2021).

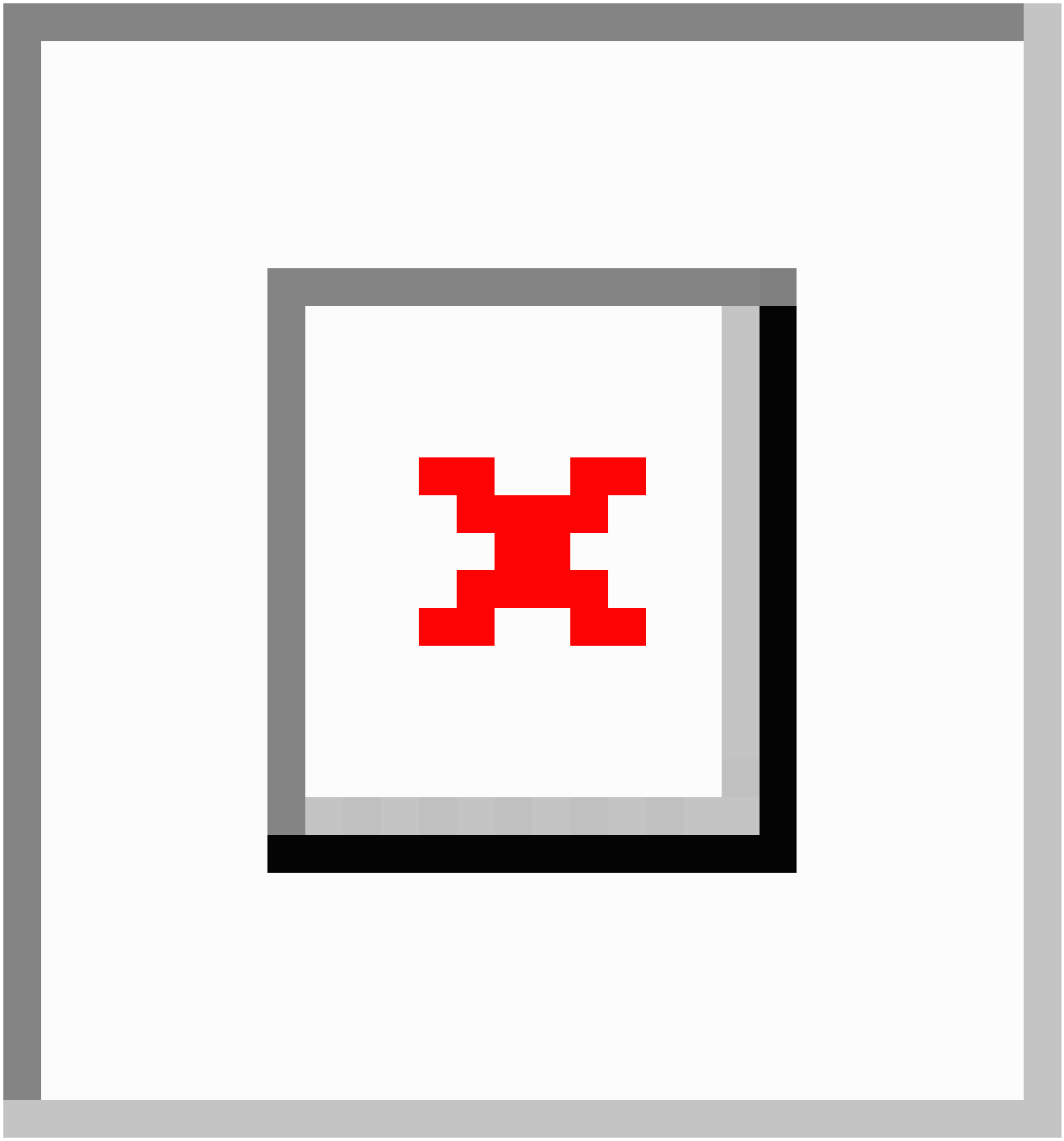
Characteristics		Female patients (n=28,109, 68.81%)	Male patients (n=18,113, 39.19%)
Age at first visit			
	Years, mean (SD)	50.46 (18.82)	48.72 (19.02)
	Years, median (range)	49.0 (0-109)	47.0 (2-98)
Number of charts/patient			
	Value, mean (SD)	1.62 (1.55)	1.64 (2.38)
	Value, median (range)	1.0 (1-63)	1.0 (1-132)
Patients with at least one hospitalization, n		2680	2504
Number of hospitalizations/patient			
	Value, mean (SD)	0.33 (2.36)	0.46 (2.35)
	Value, median (range)	0 (0-143)	0 (0-102)
Patients with at least one compulsory admission, n (%)		415 (1.48)	485 (2.68)
Duration of hospitalization			
	Days, mean (SD)	5.08 (40.89)	7.53 (62.39)
	Days, median (range)	0 (0-2661)	0 (0-4090)
First recorded mental disorder diagnosis[a], n (%)			
	Anxiety disorders	6884 (24.49)	3725 (20.57)
	Dementia and other organic disorders	2092 (7.44)	1601 (8.84)
	Depression	7648 (27.21)	3335 (18.41)
	Drug and substance use or abuse	415 (1.48)	861 (4.75)
	Eating disorders	241 (0.86)	18 (0.10)
	Intellectual disability	468 (1.66)	636 (3.51)
	Mania and bipolar disorders	713 (2.54)	460 (2.54)
	Personality disorders	1287 (4.58)	1186 (6.55)
	Schizophrenia and other nonorganic psychoses	1468 (5.22)	1515 (8.36)
	Other mental disorders	2339 (8.32)	1130 (6.24)
	No formal mental disorder diagnosis	4554 (16.20)	3646 (20.13)

[a]Mental disorder diagnoses: *International Classification of Diseases, Ninth Revision* codes 290.xx-319.xx.

Clustering Results

This analysis, which was carried out on the subset of 38,022 individuals who had at least one mental disorder diagnosis, identified 2 distinct clusters (Figure 3). One is represented by single male patients who were born in Ferrara, those who were living with parents, and those whose age at first visit was between 25 and 34 years; the other is represented by married female patients who were living with their own acquired families, those who were born outside the province of Ferrara, and those whose age at first visit was 35 to 44 years. The following sociodemographic features were similar in the two clusters: Italian nationality, individuals with a high school degree, employed individuals, and individuals who were receiving treatment in the Ferrara catchment area.

Figure 3. Results from the clustering analysis.



Discussion

Principal Findings

This study describes the process of adapting one of the longest running EHRs of public mental health care for research purposes. The FEPSY data set covers a catchment area with 342,061 inhabitants (as of 2020) and includes a total of 46,222 unique individuals who had access to mental health services over a span of 30 years (1991-2021). The FEPSY database is suitable for descriptive, predictive, and inferential analyses via conventional analysis and AI techniques, as demonstrated by the preliminary findings of the clustering analysis. To our knowledge, our database is the first of its kind in Italy. In Europe, longitudinal and prospective registries have long been in use. For example, large data sets were extracted from the Danish National Patient Registry [30] and the Danish Psychiatric Central Research Register [31]. The first data set contained data on 8,085,603 patients, which were collected from 1977 to 2012. The second data set included data on a total of 747,176 patients, which were collected from 1970 to 2010. In both cases, the register contained dates of the onset and end of any treatment, diagnoses, types of referrals, and places of treatment, thereby allowing for the possibility to perform health registry–based research [32], considering a total population of approximately 27 million.

The main finding of this study is that data that were not originally conceived for research were successfully extracted from EHR software and loaded into a new anonymized database.

This step is of foremost importance, as the data originated from an information system that was changed and updated multiple times and was not designed to allow for exploratory investigations in a structured manner. Thus, this new data set may represent the ideal setting to build, test, and refine an analytical methodology for extracting data and preparing these data for research purposes. This methodology could also be applied to other clinical data sets, such as data sets from other medical disciplines (eg, oncology), with characteristics that are similar to those of the FEPSY data set [33]. Additionally, it will be of foremost importance to validate the methodological approach and findings of upcoming research originating from the FEPSY data set by proactively seeking collaboration with other research groups, in order to enable the replication of findings from the territory of Ferrara and the use of the FEPSY data set to replicate findings from research involving other registries.

This work also allows for collaborations in terms of learning health networks, which use comparable data that originate from EHRs to support clinical decisions, improve the delivery of efficient and effective medical care, and help with the integration of research in health care [34-36].

Our results will also pave the way for an in-depth study on the use of health care resources; the results will be used to develop a system that is capable of planning the use of such resources. Such a system would optimize the use of health care resources while maintaining or possibly improving the quality of treatment. For example, in Italy, Donisi et al [37] predicted the cost of community mental health care by using clinical and sociodemographic information originating from the Psychiatric Case Register in the Verona Health District. This allowed for the linking of social deprivation to psychiatric service utilization [38] and shed a light on possible contributors to social isolation in an already vulnerable population. Our clustering analysis, which was conducted on the FEPSY data set to test its feasibility and robustness, identified 2 clusters; women appeared to access mental health services later in life and were typically married, in comparison to men. These findings were consistent with the literature [39-41] and supported the ability of the extracted data to detect known patterns, even though the results should be interpreted with caution, given the large amount of missing sociodemographic data. Clustering analyses can be useful for building prediction models and planning a department's resource allocation, as they provide relevant information on patients at presentation and on illness trajectory [42].

The process described in this paper faced 5 major challenges that we mitigated, as follows. First, the source data included several built-in structural informatic elements (so-called *tables*) that had to be screened and deleted in order to get to the core data. Second, the anonymization step was of absolute importance, and in order to both comply with privacy constraints and be able to preserve the integrity of the data, the study team decided to keep only subelements of certain data items (eg, for the birth date, only the year was kept, and for residence, only the postal code was kept). Third, in order to establish which records were correct, an iterative comparison of the FEPSY database and the local and regional database was performed by a third party who had access to PHI. Fourth, records that were deemed unreliable were excluded (eg, clinical procedures referring to nonexistent medical records). Furthermore, the data extracted from EFESO originated from different software and thus possibly generated some errors. In the FEPSY data set, these errors seem to be limited to records, and the proportion of records with errors was very low (49,854/3,861,432, 1.29%). In the end, we decided to exclude data that were collected before 1991 and duplicate patient records. Fifth, missing data challenges were also addressed, especially in the clustering process. For this purpose, the WEKA data mining tool was used.

Strengths of This Study

The quality and completeness of the collected and cleaned data, as well as the large number of records stored in the FEPSY database, resulted in the definition of a data set that is particularly suitable for automatic analysis and has appealing characteristics for research, such as a long period of data availability, great diversity in the sociodemographic factors of the patients represented, and a history of treatments and drugs administered. This could possibly represent a strong foundation for many different studies of mental illness and resource use, favoring comparisons between Italy and other countries regarding the delivery and quality of community and hospital psychiatric care [43-45]. Furthermore, the newly created database does not include sensitive information, even though this information can be retrieved by using an external supporting table created ad hoc, which could link the FEPSY data set with other data sets (eg, hospital data and tumor registries) for future research.

The novelty of this project is represented by its interdisciplinary nature (psychiatry, public health, epidemiology, sociology, mathematics, computer science, and AI), the potential versatility of the methods that can be used with the FEPSY database, and the versatility of the systems that could be created via analyses involving the FEPSY database. To our knowledge, this study is the first attempt to retrospectively build a single data set that includes more than 30 years' worth of data on mental health services in a specific area.

Such a data set would also allow for longitudinal analyses, such as those that have already been performed with the Nordic registry (a prospective registry) and, more recently, the South London and Maudsley National Health Service [46] and the Camden & Islington Research Database [47].

We believe that historical data can add value to subsequent analyses, because they allow researchers to understand how mental health services have evolved over the past decades and the extent to which phenotypical presentations of different diseases have changed over time. In light of these considerations, factors that should be taken into account are (1) potential cohort and time effects, such as historical events (eg, the Great Recession in 2008); (2) changes in legal and medical approaches to mental health; and (3) changes in the classification of mental disorders [48].

Limitations

Our findings must be interpreted in the light of some limitations. First, the sample size is limited by the geographical catchment. A larger catchment or a more densely populated region would

probably have a larger volume of treated individuals and thus have more data, which would facilitate machine learning analyses. However, we believe that even if the sample is limited by the geographical catchment, the diverse socioeconomic distribution in Ferrara is a strength that mitigates this limitation, providing insight into the possible moderator or mediator roles of socioeconomic variables that are considered social determinants of mental health. Second, another potential limitation is the missing data for some sociodemographic attributes, which may reduce the statistical power of a study [49] or affect the accuracy of machine learning algorithms [50]. In order to overcome this issue, missing values can be handled with multiple imputation methods or replaced with the mean or the mode (ie, for quantitative or qualitative data, respectively). Moreover, sociodemographic information can be drawn from external and publicly available sources, such as the Italian National Institute of Statistics [51], which includes the census of the population as well as social, economic, and environmental surveys and analyses. Lastly, there is the risk of introducing bias while building prediction models, especially when using supervised machine learning techniques, due to small sample sizes and the poor handling of missing data and overfitting [52]. With regard to the sample size, our sample appears to be sufficiently large for risk prediction analyses. Overfitting can be addressed with penalized models [53].

Future Directions

This work sets a starting point for future investigations, which can be described as follows: (1) identifying patients who have a higher severity index or chronicity level and those who require a greater use of health resources; (2) identifying and validating, by means of machine learning models, demographic, clinical, and social predictors of clinically relevant outcomes that are useful for an ad hoc programming of resources (eg, sex, gender, or social deprivation [38,39,54]); (3) further optimizing and tailoring the analysis methods, so that they can also be applied to other data sets (eg, the local mental health registries for child and adolescent neuropsychiatry and for drug addiction services); (4) interacting with international learning health networks [55-57]; and (5) linking the FEPSY data set with external data sources, such as census data, tumor registries [58], death registries, and criminal justice data [59]. As a result of the increasing digitalization of medical records, it was possible to gather years of mental health history for every patient. This will enable for the conduct of symbolic and subsymbolic analyses on time series via automatic methodologies. Classical supervised and unsupervised machine learning and deep learning techniques will be evaluated. In order to explore the relationship between sociodemographic characteristics and specific diagnostic questions (eg, the incidence and prevalence of psychosis), a supervised framework will be deployed, in which binary labels (eg, “psychosis” or “no-psychosis”) or multiple classification labels (eg, “ICD-9 diagnosis”) will be associated with the patient. The main problem to overcome will be the imbalance of the data set, that is, when there is an unequal distribution of classes in the data set. In such instances, a standard machine learning technique, such as a support vector machine or random forest [60,61], will be applied. Moreover, each patient could potentially be considered as a distinct time series by including the temporal dimension of the treatment and by applying recurrent neural networks [62,63]. By doing so, the prediction of the new onset of a disease and the subsequent use of health resources will be the focus, in order to plan and optimize health care resources.

Conclusions

The process described in this study resulted in the building of a data set that included the information of 46,222 individuals who had access to psychiatric services in the Ferrara province over the course of almost 30 years. The preliminary findings from the clustering analysis confirmed the quality of the newly established database. The process we implemented proved to be a solid method that can be replicated with similar data sets, even if they were not originally compiled for research purposes.

Acknowledgments

This work was supported by the “Fondo per l’Incentivazione alla Ricerca (FIR),” granted by the University of Ferrara in 2021 to MF, with the project titled “Intelligenza Artificiale per Predizione Diagnostica, di Carico Assistenziale, e di Esito: 40 Anni di accessi Presso i Centri di Salute Mentale della Provincia di Ferrara (AI4MentalHealth).”

Authors' Contributions

MF designed the study, wrote the first draft of the manuscript, and supervised the project. EG processed the experimental data, performed the analysis, and designed the figures. MBM, RZ, MA, GF, ID, FF, and LG were involved in planning and interpreting the results. CS, LB, and JL aided in interpreting the results. All authors discussed the results, commented on the manuscript, and approved the final version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Characteristics of the records within the tables of the Ferrara-Psychiatry (FEPSY) database.
[DOCX File, 17 KB - medinform_v11i1e45523_app1.docx]

References

1. Braveman P, Gottlieb L. The social determinants of health: it's time to consider the causes of the causes. Public Health Rep 2014;129 Suppl 2(Suppl 2):19-31. [doi: 10.1177/00333549141291S206] [Medline: 24385661]
2. World Health Organization. Social determinants of health. URL: www.who.int/health-topics/social-determinants-of-health#tab=tab_1 [accessed 2023-07-4]
3. Marmot M, Bell R. Social determinants and non-communicable diseases: time for integrated action. BMJ 2019 Jan 28;364:l251. [doi: 10.1136/bmj.l251] [Medline: 30692093]
4. Thieme A, Belgrave D, Doherty G. Machine learning in mental health: a systematic review of the HCI literature to support the development of effective and implementable ML systems. ACM Trans Comput Hum Interact 2020 Aug 17;27(5):1-53. [doi: 10.1145/3398069]
5. Mehta N, Pandit A, Shukla S. Transforming healthcare with big data analytics and artificial intelligence: A systematic mapping study. J Biomed Inform 2019 Dec;100:103311. [doi: 10.1016/j.jbi.2019.103311] [Medline: 31629922]
6. Ngiam KY, Khor IW. Big data and machine learning algorithms for health-care delivery. Lancet Oncol 2019 May;20(5):e262-e273. [doi: 10.1016/S1470-2045(19)30149-4] [Medline: 31044724]
7. Mechelli A, Vieira S, editors. Machine Learning: Methods and Applications to Brain Disorders, 1st Edition. San Deigo, CA: Elsevier; 2019. ISBN.
8. Graham S, Depp C, Lee EE, Nebeker C, Tu X, Kim HC, et al. Artificial intelligence for mental health and mental illnesses: an overview. Curr Psychiatry Rep 2019 Nov 7;21(11):116. [doi: 10.1007/s11920-019-1094-0] [Medline: 31701320]
9. Hughes MC, Pradier MF, Ross AS, McCoy THJ, Perlis RH, Doshi-Velez F. Assessment of a prediction model for antidepressant treatment stability using supervised topic models. JAMA Netw Open 2020 May 1;3(5):e205308. [doi: 10.1001/jamanetworkopen.2020.5308] [Medline: 32432711]
10. Xu Z, Wang F, Adekkanattu P, Bose B, Vekaria V, Brandt P, et al. Subphenotyping depression using machine learning and electronic health records. Learn Health Syst 2020 Aug 3;4(4):e10241. [doi: 10.1002/lrh2.10241] [Medline: 33083540]
11. Pradier MF, Hughes MC, McCoy THJ, Barroilhet SA, Doshi-Velez F, Perlis RH. Predicting change in diagnosis from major depression to bipolar disorder after antidepressant initiation. Neuropsychopharmacology 2021 Jan;46(2):455-461. [doi: 10.1038/s41386-020-00838-x] [Medline: 32927464]
12. Perlis RH, Iosifescu DV, Castro VM, Murphy SN, Gainer VS, Minnier J, et al. Using electronic medical records to enable large-scale studies in psychiatry: treatment resistant depression as a model. Psychol Med 2012 Jan;42(1):41-50. [doi: 10.1017/S0033291711000997] [Medline: 21682950]
13. van der Lei J. Use and abuse of computer-stored medical records. Methods Inf Med 1991 Apr;30(2):79-80. [Medline: 1857252]]
14. Park HA, Jung H, On J, Park SK, Kang H. Digital epidemiology: use of digital data collected for non-epidemiological purposes in epidemiological studies. Healthc Inform Res 2018 Oct;24(4):253-262. [doi: 10.4258/hir.2018.24.4.253] [Medline: 30443413]
15. Gazzetta Ufficiale. Legge 22 Marzo 2019, n 29: Istituzione e disciplina della Rete nazionale dei registri dei tumori e dei sistemi di sorveglianza e del referto epidemiologico per il controllo sanitario della popolazione. 2019. URL: www.gazzettaufficiale.it/eli/id/2019/04/05/19G00036/sg [accessed 2023-07-4]
16. Ferrannini L, Ghio L, Gibertoni D, Lora A, Tibaldi G, Neri G, et al. Thirty-five years of community psychiatry in Italy. J Nerv Ment Dis 2014 Jun;202(6):432-439. [doi: 10.1097/NMD.0000000000000141] [Medline: 24821278]
17. Barbui C, Papola D, Saraceno B. Forty years without mental hospitals in Italy. Int J Ment Health Syst 2018 Jul 31;12:43. [doi: 10.1186/s13033-018-0223-1] [Medline: 30079100]
18. Mezzina R. Forty years of the law 180: the aspirations of a great reform, its successes and continuing need. Epidemiol Psychiatr Sci 2018 Aug;27(4):336-345. [doi: 10.1017/S2045796018000070] [Medline: 29506591]
19. Gazzetta Ufficiale. Decreto del Presidente Della Repubblica 1 Novembre 1999: Approvazione del Progetto Obiettivo “Tutela salute Mentale 1998-2000”. 1999. URL: www.gazzettaufficiale.it/atto/serie_generale/caricaDettaglioAtto/originario?atto.dataPubblicazioneGazzetta=1999-11-22&atto.codiceRedazionale=099A9917&elenco30giorni=false [accessed 2023-07-4]
20. Ferrara M, Tedeschini E, Baccari F, Musella V, Vacca F, Mazzi F, et al. Early intervention service for first episode psychosis in Modena, Northern Italy: the first hundred cases. Early Interv Psychiatry 2019 Aug;13(4):1011-1017. [doi: 10.1111/eip.12788] [Medline: 30672134]
21. Murri MB, Bertelli R, Carozza P, Berardi L, Cantarelli L, Croce E, et al. First-episode psychosis in the Ferrara Mental Health Department: incidence and clinical course within the first 2 years. Early Interv Psychiatry 2021 Dec;15(6):1738-1748. [doi: 10.1111/eip.13095] [Medline: 33264815]
22. Denney MJ, Long DM, Armistead MG, Anderson JL, Conway BN. Validating the extract, transform, load process used to populate a large clinical research database. Int J Med Inform 2016 Oct;94:271-274. [doi: 10.1016/j.ijmedinf.2016.07.009] [Medline: 27506144]
23. Ralambondrainy H. A conceptual version of the K-means algorithm. Pattern Recognit Lett 1995 Nov;16(11):1147-1157. [doi: 10.1016/0167-8655(95)00075-R]

24. Witten IH, Frank E, Hall MA, Pal CJ. Data Mining: Practical Machine Learning Tools and Techniques, Fourth Edition. Burlington, MA: Morgan Kaufmann Publishers Inc; 2016.
25. World Health Organization. International Classification of Diseases Ninth Revision: Basic Tabulation List With Alphabetic Index. Geneva, Switzerland: World Health Organization; 1978.
26. The Pandas development team. pandas-dev/Pandas: Pandas 1.3.4. 2021. URL: zenodo.org/record/5574486/export/hx [accessed 2023-07-6]
27. McKinney W. Data structures for statistical computing in Python. June 28 to July 3, 2010 Presented at: Presented at Proceedings of the 9th Python in Science Conference (SciPy 2010); Austin, Texas p. 56-61. [doi: 10.25080/Majora-92bf1922-00a]
28. Leach PJ, Salz R, Mealling MH. A universally unique identifier (UUID) URN namespace. 2005. URL: dl.acm.org/doi/pdf/10.17487/RFC4122 [accessed 2023-07-6]
29. Van Rossum G, Drake FL. Python 3 Reference Manual. Scotts Valley, CA: CreateSpace; 2009.
30. Schmidt M, Schmidt SAJ, Sandegaard JL, Ehrenstein V, Pedersen L, Sørensen HT. The Danish National Patient Registry: a review of content, data quality, and research potential. Clin Epidemiol 2015 Nov 17;7:449-490. [doi: 10.2147/CLEP.S91125] [Medline: 26604824]
31. Mors O, Perto GP, Mortensen PB. The Danish Psychiatric Central Research Register. Scand J Public Health 2011 Jul;39(7 Suppl):54-57. [doi: 10.1177/1403494810395825] [Medline: 21775352]
32. Laugesen K, Ludvigsson JF, Schmidt M, Gissler M, Valdimarsdottir UA, Lunde A, et al. Nordic health registry-based research: A review of health care systems and key registries. Clin Epidemiol 2021 Jul 19;13:533-554. [doi: 10.2147/CLEP.S314959] [Medline: 34321928]
33. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. Reply. N Engl J Med 2019 Jun 27;380(26):2589-2590. [doi: 10.1056/NEJMc1906060] [Medline: 31242381]
34. Vargas N, Lebrun-Harris LA, Weinberg J, Dievler A, Felix KL. Qualitative perspective on the learning health system: how the Community Health Applied Research Network paved the way for research in safety-net settings. Prog Community Health Partnersh 2018;12(3):329-339. [doi: 10.1353/cpr.2018.0057] [Medline: 30581176]
35. Squires JE, Logan B, Lorts A, Haskell H, Sisaithong K, Pillari T, et al. A learning health network for pediatric liver transplantation: inaugural meeting report from the Starzl Network for Excellence in Pediatric Transplantation. Pediatr Transplant 2019 Sep;23(6):e13528. [doi: 10.1111/petr.13528] [Medline: 31328841]
36. Marsolo K, Margolis PA, Forrest CB, Colletti RB, Hutton JJ. A digital architecture for a network-based learning health system: integrating chronic care management, quality improvement, and research. EGEMS (Wash DC) 2015 Aug 17;3(1):1168. [doi: 10.13063/2327-9214.1168] [Medline: 26357665]
37. Donisi V, Jones J, Pertile R, Salazzari D, Grigoletti L, Tansella M, et al. The difficult task of predicting the costs of community-based mental health care. A comprehensive case register study. Epidemiol Psychiatr Sci 2011 Sep;20(3):245-256. [doi: 10.1017/s2045796011000473] [Medline: 21922967]
38. Thornicroft G, Bisoffi G, De Salvia D, Tansella M. Urban-rural differences in the associations between social deprivation and psychiatric service utilization in schizophrenia and all diagnoses: a case-register study in Northern Italy. Psychol Med 1993 May;23(2):487-496. [doi: 10.1017/s0033291700028579] [Medline: 8332662]
39. Ferrara M, Srihari VH. Early intervention for psychosis in the United States: Tailoring services to improve care for women. Psychiatr Serv 2021 Jan 1;72(1):5-6. [doi: 10.1176/appi.ps.202000205] [Medline: 32966169]
40. Ministero della Salute. Piano per l'applicazione e la diffusione della Medicina di Genere. 2019. URL: www.salute.gov.it/portale/donna/dettaglioPubblicazioniDonna.jsp?id=2860&lingua=italiano [accessed 2023-07-6]
41. Astbury J. Gender disparities in mental health. May 14-22, 2001 Presented at: Presented at Fifty-fourth World Health Assembly; Geneva, Switzerland..
42. Meisner J, Rasmussen S, Benros ME. Towards precision psychiatry utilizing large-scale multimodal data paving the way for improved prevention and treatment of mental disorders. Neuroscience Applied 2023;2:101017. [doi: 10.1016/j.nsa.2022.101017]
43. Guaiana G, O'Reilly R, Grassi L. A comparison of inpatient adult psychiatric services in Italy and Canada. Community Ment Health J 2019 Jan;55(1):51-56. [doi: 10.1007/s10597-018-0283-3] [Medline: 29725879]
44. Martinelli A, Iozzino L, Ruggeri M, Marston L, Killaspy H. Mental health supported accommodation services in England and in Italy: a comparison. Soc Psychiatry Psychiatr Epidemiol 2019 Nov;54(11):1419-1427. [doi: 10.1007/s00127-019-01723-9] [Medline: 31055632]
45. Bird V, Miglietta E, Giacco D, Bauer M, Greenberg L, Lorant V, et al. Factors associated with satisfaction of inpatient psychiatric care: a cross country comparison. Psychol Med 2020 Jan;50(2):284-292. [doi: 10.1017/S0033291719000011] [Medline: 30696510]
46. Perera G, Broadbent M, Callard F, Chang CK, Downs J, Dutta R, et al. Cohort profile of the South London and Maudsley NHS Foundation Trust Biomedical Research Centre (slam BRC) Case Register: current status and recent enhancement of an electronic mental health record-derived data resource. BMJ Open 2016 Mar 1;6(3):e008721. [doi: 10.1136/bmjopen-2015-008721] [Medline: 26932138]

47. Werbeloff N, Osborn DPJ, Patel R, Taylor M, Stewart R, Broadbent M, et al. The Camden & Islington Research Database: using electronic mental health records for research. PLoS One 2018 Jan 29;13(1):e0190703. [doi: 10.1371/journal.pone.0190703] [Medline: 29377897]
48. Andreoli RRV, Cassano GB. DSM-IV-TR. Manuale diagnostico e statistico dei disturbi mentali. Text revision. ICD-10/ICD-9-CM. Classificazione parallela. Amsterdam, Netherlands: Elsevier; 2007.
49. Kang H. The prevention and handling of the missing data. Korean J Anesthesiol 2013 May;64(5):402-406. [doi: 10.4097/kjae.2013.64.5.402] [Medline: 23741561]
50. Makaba T, Dogo E. A comparison of strategies for missing values in data on machine learning classification Algorithms. November 21-22, 2019 Presented at: Presented at 2019 International Multidisciplinary Information Technology and Engineering Conference (IMITEC); Vanderbijlpark, South Africa p. 1-7. [doi: 10.1109/IMITEC45504.2019.9015889]
51. Istat | Istituto Nazionale di Statistica. URL: www.istat.it [accessed 2020-04-13]
52. Navarro CLA, Damen JAA, Takada T, Nijman SWJ, Dhiman P, Ma J, et al. Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. BMJ 2021 Oct 20;375:n2281. [doi: 10.1136/bmj.n2281]
53. Moons KGM, Donders ART, Steyerberg EW, Harrell FE. Penalized maximum likelihood estimation to directly adjust diagnostic and prognostic prediction models for overoptimism: a clinical example. J Clin Epidemiol 2004 Dec;57(12):1262-1270. [doi: 10.1016/j.jclinepi.2004.01.020] [Medline: 15617952]
54. Franceschini A, Fattore L. Gender-specific approach in psychiatric diseases: because sex matters. Eur J Pharmacol 2021 Apr 5;896:173895. [doi: 10.1016/j.ejphar.2021.173895] [Medline: 33508283]
55. Heinssen RK, National Institute of Mental Health. Early Psychosis Intervention Network (EPINET): a learning healthcare system for early serious mental illness. 2015. URL: www.nimh.nih.gov/funding/grant-writing-and-application-process/concept-clearances/2015/early-psychosis-intervention-network-epinet-a-learning-healthcare-system-for-early-serious-mental-illness [accessed 2023-07-6]
56. STEP Learning Collaborative. URL: www.ctearlypsychosisnetwork.org/ [accessed 2023-07-21]
57. Srihari VH, Ferrara M, Li F, Kline E, Gülöksüz S, Pollard JM, et al. Reducing the duration of untreated psychosis (DUP) in a US community: a quasi-experimental trial. Schizophr Bull Open 2022 Jan 4;3(1):sgab057. [doi: 10.1093/schizbullopen/sgab057] [Medline: 35295656]
58. Grassi L, Stivanello E, Murri MB, Perlangeli V, Pandolfi P, Carnevali F, et al. Mortality from cancer in people with severe mental disorders in Emilia Romagna Region, Italy. Psychooncology 2021 Dec;30(12):2039-2051. [doi: 10.1002/pon.5805] [Medline: 34499790]
59. Pollard JM, Ferrara M, Lin IH, Kucukgoncu S, Wasser T, Li F, et al. Analysis of early intervention services on adult judicial outcomes. JAMA Psychiatry 2020 Aug 1;77(8):871-872. [doi: 10.1001/jamapsychiatry.2020.0448] [Medline: 32320010]
60. Ferrara M, Franchini G, Funaro M, Cutroni M, Valier B, Toffanin T, et al. Machine learning and non-affective psychosis: identification, differential diagnosis, and treatment. Curr Psychiatry Rep 2022 Dec;24(12):925-936. [doi: 10.1007/s11920-022-01399-0] [Medline: 36399236]
61. Ferrara M, Franchini G, Funaro M, Murri MB, Toffanin T, Zerbinati L, et al. Machine learning for mental health: focus on affective and non-affective psychosis. In: Sousa MJ, Pani S, dal Mas F, Sousa S, editors. Incorporating AI Technology in the Service Sector: Innovations in Creating Knowledge, Improving Efficiency, and Elevating Quality of Life. Palm Bay, FL: Apple Academic Press Inc; 2023.
62. Lipton ZC, Kale DC, Elkan C, Wetzel R. Learning to diagnose with LSTM recurrent neural networks. arXiv. Preprint posted online on November 11, 2015. . [doi: 1511.03677]
63. Che Z, Purushotham S, Cho K, Sontag D, Liu Y. Recurrent neural networks for multivariate time series with missing values. Sci Rep 2018 Apr 17;8(1):6085. [doi: 10.1038/s41598-018-24271-9] [Medline: 29666385]

Abbreviations

AI: artificial intelligence
EHR: electronic health record
FEPSY: Ferrara-Psychiatry
***ICD-9*:** *International Classification of Diseases, Ninth Revision*
PHI: protected health information
SIPER: Sistema Informativo Psichiatrico dell'Emilia-Romagna
UUID: universally unique identifier
WEKA: Waikato Environment for Knowledge Analysis

Edited by C Lovis; submitted 05.01.23; peer-reviewed by N Mehta; revised version received 04.05.23; accepted 01.06.23; published 09.08.23.

Please cite as:
Ferrara M, Gentili E, Belvederi Murri M, Zese R, Alberti M, Franchini G, Domenicano I, Folesani F, Sorio C, Benini L, Carozza P, Little J, Grassi L
Establishment of a Public Mental Health Database for Research Purposes in the Ferrara Province: Development and Preliminary Evaluation Study
JMIR Med Inform 2023;11:e45523
URL: https://medinform.jmir.org/2023/1/e45523
doi:10.2196/45523

Original Paper

The Journey of Data Within a Global Data Sharing Initiative: A Federated 3-Layer Data Analysis Pipeline to Scale Up Multiple Sclerosis Research

Ashkan Pirmani[1,2,3,4], MSc; Edward De Brouwer[1], PhD; Lotte Geys[2,3,4], PhD; Tina Parciak[2,3,4], MSc; Yves Moreau[1*], Prof Dr; Liesbet M Peeters[2,3,4*], Prof Dr

[1]ESAT, STADIUS, KU Leuven, Leuven, Belgium

[2]Biomedical Research Institute, Hasselt University, Diepenbeek, Belgium

[3]Data Science Institute, Hasselt University, Diepenbeek, Belgium

[4]University Multiple Sclerosis Center, Hasselt University, Diepenbeek, Belgium

[*]these authors contributed equally

Corresponding Author:
Liesbet M Peeters, Prof Dr
Biomedical Research Institute
Hasselt University
Agoralaan, Building C
Diepenbeek, 3590
Belgium
Phone: 32 11 26 92 05
Email: liesbet.peeters@uhasselt.be

Abstract

Background: Investigating low-prevalence diseases such as multiple sclerosis is challenging because of the rather small number of individuals affected by this disease and the scattering of real-world data across numerous data sources. These obstacles impair data integration, standardization, and analysis, which negatively impact the generation of significant meaningful clinical evidence.

Objective: This study aims to present a comprehensive, research question–agnostic, multistakeholder-driven end-to-end data analysis pipeline that accommodates 3 prevalent data-sharing streams: individual data sharing, core data set sharing, and federated model sharing.

Methods: A demand-driven methodology is employed for standardization, followed by 3 streams of data acquisition, a data quality enhancement process, a data integration procedure, and a concluding analysis stage to fulfill real-world data-sharing requirements. This pipeline's effectiveness was demonstrated through its successful implementation in the COVID-19 and multiple sclerosis global data sharing initiative.

Results: The global data sharing initiative yielded multiple scientific publications and provided extensive worldwide guidance for the community with multiple sclerosis. The pipeline facilitated gathering pertinent data from various sources, accommodating distinct sharing streams and assimilating them into a unified data set for subsequent statistical analysis or secure data examination. This pipeline contributed to the assembly of the largest data set of people with multiple sclerosis infected with COVID-19.

Conclusions: The proposed data analysis pipeline exemplifies the potential of global stakeholder collaboration and underlines the significance of evidence-based decision-making. It serves as a paradigm for how data sharing initiatives can propel advancements in health care, emphasizing its adaptability and capacity to address diverse research inquiries.

(JMIR Med Inform 2023;11:e48030) doi:10.2196/48030

KEYWORDS

data analysis pipeline; federated model sharing; real-world data; evidence-based decision-making; end-to-end pipeline; multiple sclerosis; data analysis; pipeline; data science; federated; neurology; brain; spine; spinal nervous system; neuroscience; data sharing; rare; low prevalence

Introduction

Chronic diseases such as multiple sclerosis (MS) [1] present significant obstacles for research, primarily because of their limited prevalence, resulting in smaller study populations [2]. The scarcity of the affected individuals is reinforced when considering the dispersion of real-world data (RWD) across diverse repositories. This scarce RWD, sourced during routine clinical care [3,4], further coupled with heterogeneity in formats, quality standards, and regulatory guidelines, make the comprehensive collection and extraction of meaningful clinical insights even more challenging [5,6].

Despite these challenges, well-managed RWD have the potential to reveal significant patterns concerning diseases, patient experiences, and treatment outcomes [7,8]. For instance, during the early stages of the COVID-19 pandemic, innovative data acquisition strategies overcame data scarcity, unlocking the potential of RWD for meaningful analysis [9-11]. These specific instances underline the broader concern: the RWD landscape is rife with challenges that are often understated.

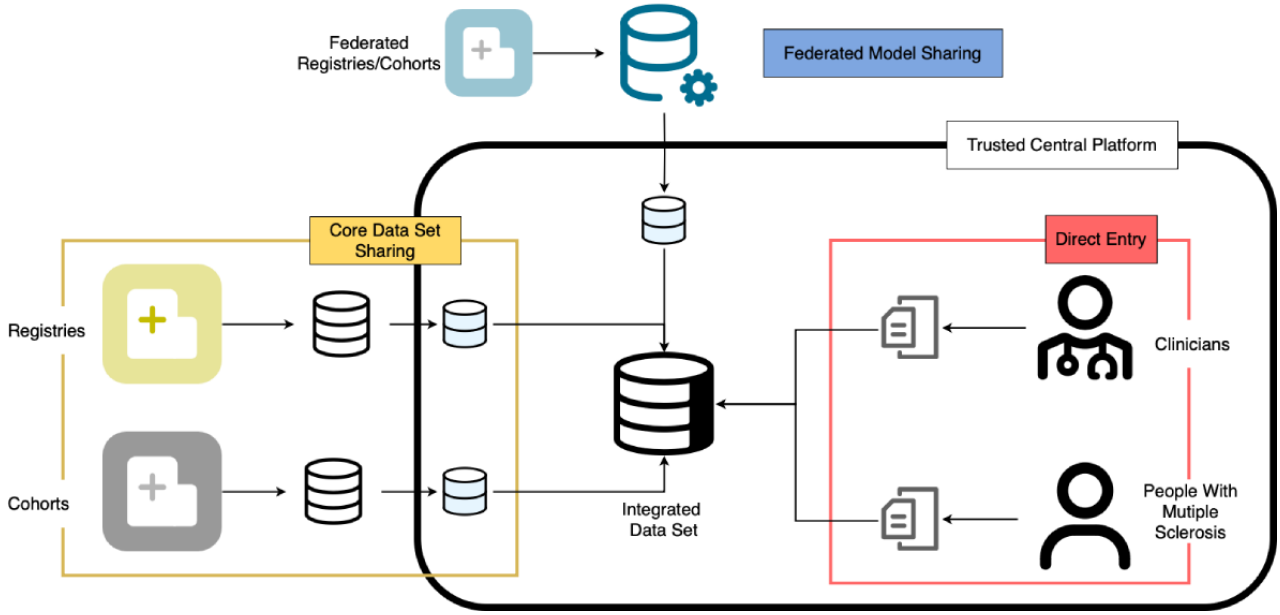
Current literature tends to oversimplify the intricate processes involved in managing RWD. These include standardization, acquisition, quality enhancement, integration, storage, governance, visualization, and eventual analysis and interpretation. Although these facets are crucial, they are often treated as isolated components rather than integral parts of an interconnected system, with certain areas occasionally overlooked [5].

Recent studies on COVID-19 bring this gap into sharper focus. Khalid and colleagues [12] focused on building analytical models by using observational health data through machine learning but did not fully emphasize the vital aspect of data acquisition in the pipeline management. By contrast, Nishimwe and colleagues [13] concentrated on data integration, gathering data from various hospitals, but did not delve into thorough in-depth data analysis. A study by Junior and colleagues [14] aimed to cover the whole data analytics pipeline but primarily focused on standardizing data from 2 different countries, giving less attention to crucial parts of RWD management, such as data acquisition, preprocessing, quality enhancement, and analysis. This fragmented focus points to the need for a more comprehensive strategy that neither compromises nor overlooks any part of the RWD management process. The absence of a holistic framework, coupled with the growing diversity and volume of RWD sources, intensifies the challenges in health care data sharing and the conversion of RWD into actionable evidence, underscoring the need for standardized management [6].

In light of these challenges, the global data sharing initiative (GDSI) emerges as an exemplary solution that addresses multiple facets of RWD management, specifically in the context of COVID-19 and MS. Prompted by the urgent need to understand COVID-19's effects on people with MS, GDSI was launched [15]. By integrating data from over 80 countries, GDSI generated globally relevant insights [7,16-19]. This large-scale effort resulted in the formation of the most extensive international cohort of COVID-19 cases among people with MS. In addition to enriching our understanding of the COVID-19 and MS interaction, GDSI showcased the enormous potential of large-scale international collaboration. Furthermore, the initiative set a methodological standard in global health research by introducing a data analysis pipeline with applications beyond MS.

This paper delves deep into GDSI's comprehensive RWD analysis pipeline, offering an end-to-end approach that spans from introducing a data dictionary to meticulous data acquisition, and ultimately, to deriving insightful clinical interpretations. One distinguishing aspect of our study lies in the pragmatic execution of this intricate end-to-end analysis pipeline. As depicted in Figure 1, we have implemented a hybrid 3-layer data acquisition architecture—all in strict compliance with the legal and ethical standards that govern data collection and dissemination. Designed for versatility and inclusivity, this architecture aimed to capture every data point possible. Concurrently, an astute approach to data integration was used, whereby these diverse data streams were seamlessly unified. This robust unified data set was then readied for further analysis and exploration.

Figure 1. The global data sharing initiative data streams detailing the initiative's inclusive approach through a hybrid 3-layer data acquisition architecture: (1) direct entry, where individuals upload their data via a web-based form; (2) core data set sharing, where registries upload patient-level data to the central platform under signed data transfer agreements and ethics approvals; and (3) federated model sharing, allowing registries with restrictive policies to participate without directly submitting patient-level data to the central platform.

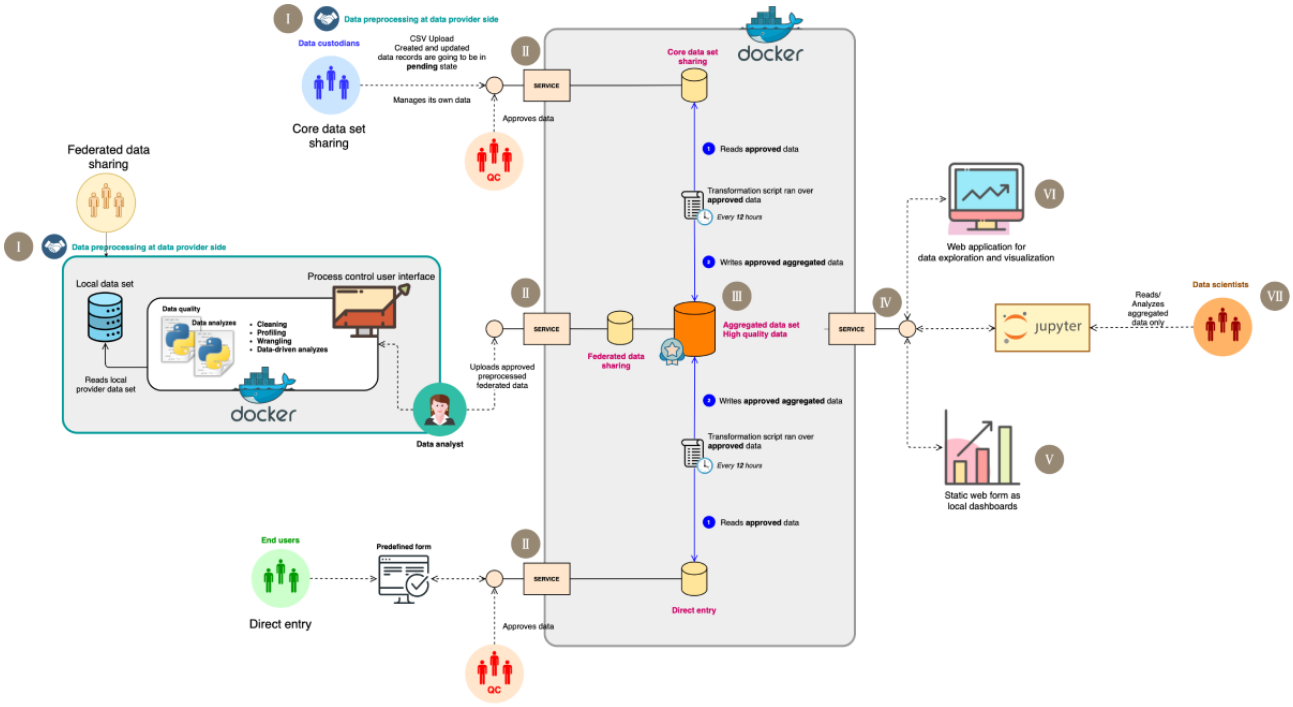


Methods

Overview of GDSI's Data Analysis Pipeline

The robustness and scale of GDSI's endeavor were mirrored by its foundational approach. As depicted in Figure 2, GDSI's RWD analysis pipeline provided the essential framework for comprehensive data management and analysis. Centralized around a core platform, this pipeline progressed through 5 key stages (1) introducing a specialized data dictionary to standardize the data; (2) data acquisition, which details the methods used to gather the data; (3) an integral step for enhancing data quality; (4) data integration, responsible for aggregating various sources; and (5) the final stage, where the consolidated data are analyzed.

Figure 2. The global data sharing initiative's end-to-end real-world data analysis pipeline. Step I illustrates the standardization process, which serves as the foundation of this architecture. In this phase, data custodians are requested to map their data to the "COVID-19 in multiple sclerosis core data set" (here referred to as the data dictionary). This process applies only to the core data set and federated model sharing registries, as direct entry is already embedded with a data dictionary via the web form. Step II involves the data acquisition pipeline, featuring distinct levels of data acquisition that depend on the data holder's willingness and internal policies, all conducted in line with ethical and legal standards. Direct entry, core data set sharing, and federated model sharing constitute the 3 data stream levels. The first 2 levels interact directly with a central platform where the core dataset is shared as static files, in this instance, Comma Separated Values (CSV), whereas federated registries necessitate additional steps before submitting outcomes to the central platform. To incorporate federated registries into the pipeline, predefined queries are dispatched alongside Docker containers to the local side of the registries. The results of these containers are then shipped back to the central platform. In step III, data from different data holders are stored in separate layers to facilitate the next data integration process. Data integration, the subsequent step in the pipeline, entails consolidating data from distinct layers into a comprehensive data set. Step IV emphasizes the utilization of the integrated data set for further data exploration and analysis. Step V highlights the local dashboard, which serves as a quality check, enabling data providers to give feedback on their uploaded data as an additional sanity check. Step VI underscores the online dashboard that has been fed by the integrated data set, utilized by the taskforce during the development of the research questions to ascertain the feasibility of the study and to monitor the data being collected. In step VII, a Jupyter Notebook is provided to the data analysis team, securely connected to the integrated data set, facilitating statistical analysis.



Ethics Approval

This study received ethics approval from the ethics committee of Hasselt University (approval CME2020/025). For an in-depth discussion concerning ethical authorization, kindly refer to Simpson-Yap et al [17].

Data Dictionary

A data dictionary serves as a guide that details the attributes of components within an information database, ensuring consistent terminology [20-22]. In the context of GDSI, this tool has proven invaluable for mitigating challenges posed by diverse languages and structures. A task force of domain experts, including epidemiologists, neurologists, and data scientists, reached a consensus on establishing the "COVID-19 in MS Core Data set" data dictionary. This guide served as a keystone for harmonizing data from various sources. To tackle interoperability, data custodians used it as a reference, enabling them to standardize their data sets and streamline the extract-transform-load process. A detailed overview of the variables employed in GDSI is provided by Simpson-Yap and colleagues [7], and a full list is accessible via the GitHub repository [23] and presented in Table S1 of Multimedia Appendix 1.

Data Acquisition

Recognizing the value of diverse data sources for research outcomes [8,24], GDSI developed a hybrid data acquisition architecture. This framework consists of 3 distinct data sharing streams: direct entry, core data set sharing, and federated model sharing. Each stream is designed to accommodate specific data environments, ensuring a comprehensive and multifaceted collection approach. The primary distinguishing factor among these data sharing streams was the extent of willingness to share clinical records with the central server. In practice, confining data collection to a singular stream would drastically reduce the data volume, making the transition of all contributors to 1 mode unattainable. GDSI's strength was rooted in its adaptability, effortlessly accommodating these 3 sharing streams and fusing them into a unified data set.

Data Sharing Streams

Direct Entry

This stream prioritized direct engagement with both clinicians and patients. Patients provided their clinical records through structured questionnaires, while clinicians offered their observations after acquiring the necessary permissions. A unique characteristic of this stream was its rapid data entry mechanism.

The predefined structure, designed to align with a condensed version of the data dictionary, ensured smooth data integration. Data were submitted via a web-based form on the centralized platform. Importantly, this form upheld patient privacy, excluding specific identifiers and enforcing stringent measures against website cookies and trackers.

Core Data Set Sharing

Adhering to the conventional approach for clinical data sharing, data providers contributed a subset of their data set to the central platform. Although this mechanism excelled in handling extensive data, it grappled with challenges related to data collaboration agreements and complex regulatory stipulations. The heterogeneity in the data format further compounded these challenges. However, the architecture of core data set sharing was designed to necessitate the schema of the data dictionary. As a result, data custodians needed to standardize their data format according to the data dictionary to upload their data using this stream. Upon achieving this congruence, custodians used the central platform's interface—a secure bridge that connected the local infrastructure of the data partners to the main platform—for data upload. For enhanced data security, once uploaded, data extraction is restricted. Additional security measures such as user activity monitoring and stringent access policies were further implemented, ensuring that registry members can only view their specific records, thus preserving data confidentiality. As the pandemic progressed, the registries were periodically invited to contribute their core data sets to the central platform.

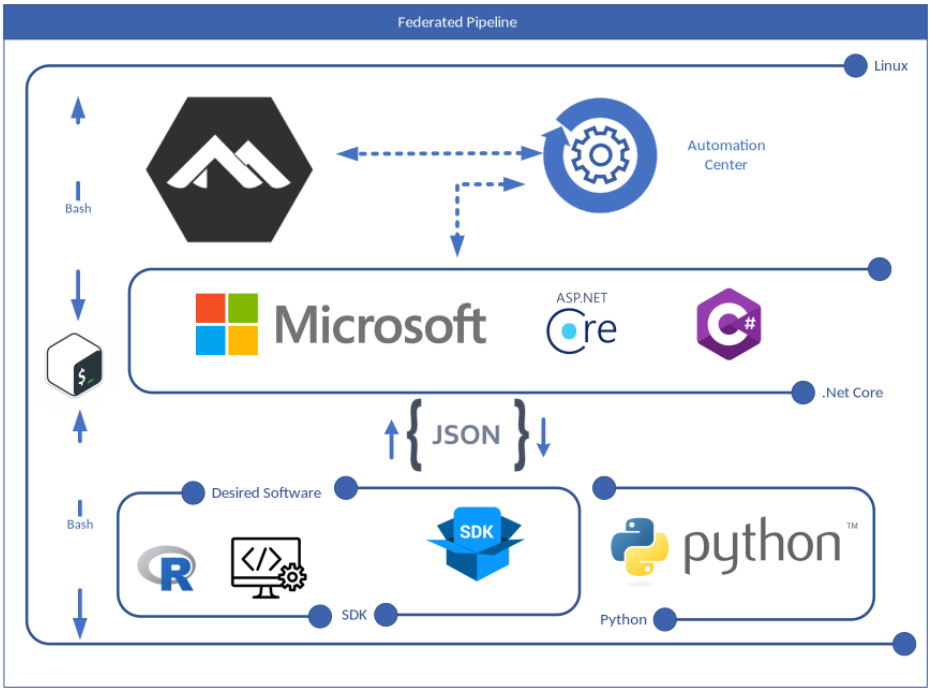
Federated Model Sharing

Addressing challenges such as strict internal policies that deterred or hindered some registries from sharing clinical records with the central platform, the federated model sharing was introduced. This decentralized solution brings a pivotal shift to regular data sharing streams. Central to this model is the principle of querying data directly at its source, thus eliminating the need to transfer patient-level data. Instead of navigating the nuances of individual patient data, this strategy consolidates multivariable categories into aggregated "buckets." These buckets are grouped categories where similar data are combined together rather than stored separately. By adopting this approach, potential risks linked to transferring patient-level data are mitigated, and the complexities tied to strict data-sharing agreements are streamlined. A detailed examination of the buckets computation methodology can be viewed in Table S2 of Multimedia Appendix 1 and within the associated GitHub repository [25].

Despite its advantages, the federated model sharing stream introduces its own challenges, especially when remote query executions result in inconsistencies across diverse systems. To compute the buckets, scripts were run locally using Docker [26] containers. Docker containers are self-contained software environments that promote standardization, which helps alleviate the typical technical challenges in such processes. These containers, referred to here as the federated pipelines, were deployed on each registry's infrastructure and were mounted with data that had already been standardized and aligned with the data dictionary, facilitating seamless execution. After computing these buckets, they were transferred to the central platform. Multiple versions of these Docker containers were used to distribute scripts across the federated model sharing registries.

The architecture of the most recent federated pipeline is presented in Figure 3 [27]. Associated resources, including a demonstrative video walk-through, operational scripts, and the Docker image, can be found in [28-31]. Furthermore, the entire source code has been made publicly accessible on GitHub [32]. This provides a thorough toolkit for those interested in understanding, replicating, or refining the framework of the federated pipeline. A comprehensive analysis of the various iterations of the federated pipeline is presented in Multimedia Appendix 2 [33].

Figure 3. The latest federated pipeline. This is a container composed of 3 primary components. The first component is the base image, which forms the bedrock of the infrastructure. This base image uses Alpine Linux as its underlying operating system, which allows the container to be fine-tuned with other software development kits for further refinements and functionalities [27]. The remaining 2 components, the backend and frontend, are constructed on top of this base image. The backend consists of a suite of Python scripts, which are tasked with data quality assessment, enhancement, cleaning, and analysis. These scripts collaboratively process the incoming mapped data, preparing it for subsequent analysis. By contrast, the frontend was crafted using Microsoft's ASP.NET Core framework and the C# programming language. Within this pipeline, there is a customizable automation center module. This module can be adapted to meet the specific needs and requests of data partners. It also integrates Crontab, a tool that automates predefined tasks and outlines complex pipelines for execution at various intervals. The automation center module also links the container to the GitHub and Docker Hub version control systems. This connection ensures the use of the most recent scripts and codes published by data analysts. SDK: software development kit.



Data Quality Assessment and Enhancement

The integrity of the acquired data set was upheld through a rigorous data enhancement and quality evaluation process, which was integrated seamlessly into the central platform. In this process, each data variable was scrutinized against a binary criterion: PASS or FAIL. If a specific data point met the pre-established benchmarks of quality and precision, it was accorded a "PASS;" otherwise, it was categorized under "FAIL." The input format for direct entry eliminated the need for additional quality checks, as validation was directly integrated into the web-based form. For the core data set sharing approach, uploaded data were immediately assessed for quality, and a real-time feedback mechanism alerted contributors to any issues, allowing for immediate corrective action. Conversely, within the federated model sharing approach, quality checks were conducted at the data source prior to aggregation.

The criteria are summarized in Table 1. In the PASS/FAIL column of this table, variables are flagged differently according to the data quality check. PASS is the flag for an accepted variable, FAIL is the flag for a dismissed variable, and EMPTY is the flag for each variable that is missing/null. Note that FAIL does not necessarily mean that the data get excluded; it is just that it is flagged as erroneous—it can also be adapted in some cases for analysis. FAIL means the following action: "Set the FAIL variable to missing, flag the variable, and keep the patient entry (row)." Additionally, specific rules were applied to dates in the data; more specifically, dates cannot be in the future (ie, if any date > date reporting, then the variable is flagged as FAIL) or before a person's birth date (ie, if any date YEAR < year_reporting - age, then the variable is flagged as FAIL). The COVID-19–related dates also must be later than the MS baseline dates (onset and diagnosis). A comprehensive version of this table is available on GitHub [34].

Table 1. Data quality assessment and enhancement: pass and fail criteria (some highlighted examples).[a]

Variable	Format	Interdependency	Pass and fail criteria
covid19_date_reporting	yyyy-mm-dd	None	if covid19_date_reporting < 2019, then fail, else pass
covid19_has_symptoms	single choice (yes/no)	covid19_sympt_fever covid19_sympt_dry_cough covid19_sympt_fatigue covid19_sympt_pain covid19_sympt_sore_throat covid19_sympt_shortness_breath covid19_sympt_nasal_congestion covid19_sympt_loss_smell_taste covid19_sympt_pneumonia	if covid19_ has_symptoms = null, then check the covid19_sympt_xx for yes if any covid19_symptxx = yes, then covid19 has symptoms = yes (for the analysis data) strict: if covid19_has_symptoms = no AND any covid19_symp_xx = yes, then fail derivation: covid19_has_symptoms is secondary to covid19_ sympt _xx if any of the single symptoms are yes, then empty(!) covid19_has_symptoms will be set to yes and vice versa (all symptoms = no, covid19_has_symptoms is set to no)
covid19_sympt_fever	single choice (yes/no)	covid19_has_symptoms	see covid19_has_symptoms
covid19_sympt_fatigue	single choice (yes/no)	covid19_has_symptoms	see covid19_has_symptoms
covid19_admission_hospital	single choice (yes/no)	None	if covid19_admission_hospital = yes AND covid19_confirmed_case = no, then fail
age_years	Integer	None	if age_years < 0 OR age_years > 110, then fail, else pass
ms_onset_date[b]	yyyy-mm-dd	ms_diagnosis_date covid19_suspected_onset	if (ms_onset_date > ms_diagnosis_date) OR (ms_onset_date > covid19_suspected_onset), then fail, else pass
edss_value[c]	Number (0.0, 10.0)	None	if edss_value < 0 OR edss_value > 10, then fail
type_dmt[d]	Single choice	None	if type_dmt = null AND type_dmt_other = null AND current_dmt = yes, then fail, else pass
has_comorbidities	single choice (yes/no)	None	if has_comorbidities = null AND any_com_xx = yes, then set has_comorbidities = yes (for analysis)

[a]67 more checks have been performed, but these checks are not presented in this table.

[b]MS: multiple sclerosis.

[c]EDSS: Expanded Disability Status Scale.

[d]DMT: disease-modifying therapy.

Data Integration

The quality-checked data acquired within each stream are stored distinctly, emphasizing the discrete nature of their origins. Consequently, the challenge emerges not just from the acquisition but notably from the critical task of integrating these separate data sets. To derive comprehensive insights, there was a paramount need to coalesce these distinct data sets into a singular unified structure. In the ensuing sections, we outline the process employed to achieve this integration and present a harmonized analytical framework.

Consider $x_i = (x_{i,1}, \ldots, x_{i,k})$ to be the list of control and response variables of patient i used in the downstream statistical analysis. N indicates the total number of patient records and K represents the number of variables of interest. For each variable type, we define a list of nonoverlapping ranges $\sum_k = \sigma_k^1, \sigma_k^2, \ldots, \sigma_k^{jk}$ that partitions the domain of each variable into distinct categories, that is, each variable $x_{i,k}$ can be categorized into a variable $y_{i,k}$ by defining $y_{i,k} = j \equiv x_{i,k} \in \sigma_k^j$ with $j \in \{1,\ldots,j_k\}$. The data extracted from the federated model sharing registries were then converted into a multivariate contingency table (S) of the patient counts for all combinations of all variables—that is, $S = \{(\sigma_0, \sigma_1,\ldots, \sigma_{K,c}): \sigma_K \in \sum_{k,c} = \sum_{i=1}^{N} I[x_{i,0} \in \sigma_0, x_{i,1} \in \sigma_1,\ldots,x_{i,k} \in \sigma_k]\}$, where $I[.]$ is the indicator function.

This set is conveniently represented as a table by considering each element of the set as a row and the columns consisting of different variable names and patient counts. This table was subsequently stored on the central platform. The same computation was performed on the direct entry and core data

set, as the raw data were available on the central platform, resulting in their specific binned count tables. Finally, all data sources were aggregated by combining their respective binned counts representation *S*. The aggregation was performed by adding the patient counts of each data source for each subgroup of variables. Then, the aggregate set was expanded into a more extended table by repeating each row several times equal to the patient count of that specific row, which resulted in a table $X \in R^{N \times K}$, with *K* as the number of variables used in the analysis and *N* as the number of patients.

Data Analysis

Following the careful integration of a diverse data set, the pivotal challenge lay in deriving actionable insights. The GDSI analysis pipeline was uniquely engineered to remain agnostic to specific clinical research inquiries. Its versatile design permitted any statistical analysis to be executed based on the variables outlined in the data integration table. The efficacy of this approach is exemplified by its application to various research questions, as highlighted in [7,16,17,19]. The analytical approach implemented by Simpson-Yap and colleagues [7] was adopted for the purpose of this paper. A multilevel mixed-effects logistic regression was employed to analyze the aggregated data table. This was performed to assess the association between disease-modifying therapies (DMTs) and several outcomes, including hospitalization, intensive care unit admission, ventilation, and death, while adjusting for variables such as age, sex, MS phenotype, and disability score. The goal of this evaluation was to determine the impact of MS-specific therapies on the severity of COVID-19. This statistical model provided a fine understanding of the complex relationship between these therapies and disease outcomes.

Results

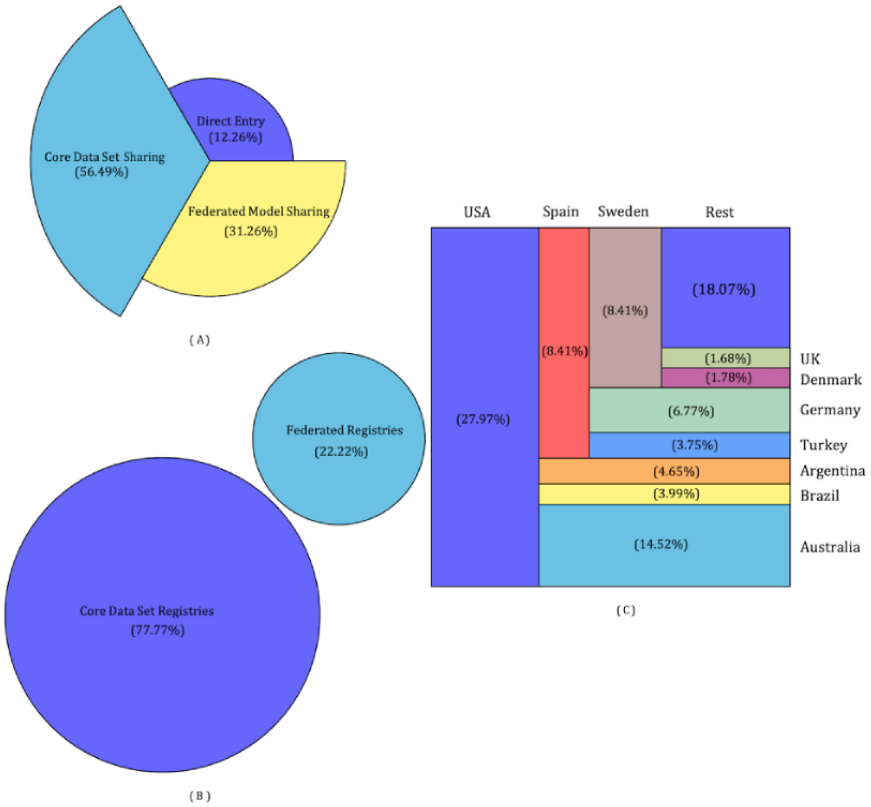
Data Acquisition

Using the pragmatic 3-layer approach of GDSI, we obtained the largest cohort of people with MS infected with COVID-19. The data were collected from 80 countries, with the top 10 contributing countries being the United States (3157/11,284, 27.97%), Australia (1639/11,284, 14.52%), Spain (949/11,284, 8.41%), Sweden (949/11,284, 8.41%), Germany (765/11,284, 6.77%), Argentina (525/11,284, 4.65%), Brazil (451/11,284, 3.99%), Turkey (424/11,284, 3.75%), Denmark (201/11,284, 1.78%), and the United Kingdom (190/11,284, 1.68%), which accounted for over 80% of the total number of records. Via direct entry, data were collected from 67 countries, with Spain contributing the largest number of records (758/1383, 54.80%), followed by the Netherlands (95/1383, 6.86%), United Kingdom (80/1383, 5.78%), United States (53/1383, 3.83%), Australia (40/1383, 2.89%), and 62 other countries (357/1383, 25.81%), resulting in a total of 1383 records. Data were collected from 18 different registries worldwide. Fourteen of these participated in core data set sharing, contributing to 6374 records. Meanwhile, 4 used the federated model sharing approach, contributing to an additional 3527 records. Table 2 enumerates these data sources. Figure 4 summarizes the number of records acquired at each stream of the data acquisition pipeline. Data acquired through direct entry have been released and are accessible through the associated PhysioNet repository [35].

Table 2. Data acquisition summary in the global data sharing initiative (N=11,284).

Method of data sharing	Values, n (%)
Direct entry	1383 (12.26)
Core data set sharing	6374 (56.49)
Federated model sharing	3527 (31.26)

Figure 4. Summary of the data acquired by implementing the 3-layer data acquisition. (A) Federated registries contribute to 31.26% (3527/11,284) of the data, while core data set sharing accounts for 56.49% (6374/11,284). (B) Only 22% (4/18) of the registries participated as federated registries. (C) A summary of the top 10 countries contributing data.



Data Analysis

Within the data analysis conducted to assess the impact of different DMTs on COVID-19 severity [7], random effects were grouped by data sources. The following variables were used: age, sex, MS phenotype, disability score, DMTs, and COVID-19 severity. Age was categorized in the following ranges: Σ_{age} = (18-50 years, 50-70 years, >70 years). Sex was binarized into male and female. MS phenotype was binarized into relapsing-remitting MS and progressive MS. The disability score was dichotomized into $\Sigma_{\text{Expanded Disability Status Scale}}$ = (0,6),(6,10). DMTs were categorized into Σ_{DMT} = (untreated, alemtuzumab, cladribine, dimethyl fumarate, fingolimod, glatiramer acetate, interferon, natalizumab, ocrelizumab, rituximab, teriflunomide, and other). COVID-19 severity was categorized into $\Sigma_{severity}$ = (hospitalization, intensive care unit admission, ventilation, death). Compared to patients using all other DMTs, those using rituximab had a higher risk of hospitalization (adjusted odds ratio [aOR] 2.76, 95% CI 1.87-4.07), intensive care unit admission (aOR 4.32, 95% CI 2.27-8.23), and artificial ventilation (aOR 6.15, 95% CI 3.09-12.27). Ocrelizumab showed similar trends for hospitalization (aOR 1.75, 95% CI 1.29-2.38) and intensive care unit admission (aOR 2.55, 95% CI 1.49-4.36) but not ventilation (aOR 1.60, 95% CI 0.82-3.14). Neither rituximab (aOR 1.72, 95% CI 0.58-5.10) nor ocrelizumab (aOR 0.73, 95% CI 0.32-1.70) were significantly associated with the risk of death. A comprehensive report of these findings can be found in Simpson-Yap et al [7].

Discussion

Insights From the GDSI Study on MS and COVID-19

The COVID-19 pandemic underscored a pressing need to understand its effect on people with MS. Recognizing the criticality of solid evidence for disease management, a global strategy involving neurologists, patients, and registries was adopted. This collaborative approach paved the way for GDSI's formation and the development of an end-to-end RWD analysis pipeline. Through this effort, GDSI emerged as the most comprehensive federated international cohort of people with MS impacted by COVID-19, becoming an invaluable resource for informed decision-making. Nevertheless, deriving conclusions from such data initiatives requires careful consideration of the inherent limitations of observational study designs. These studies provide unparalleled real-world insights, but it remains essential to situate the data within the confines of each study's specific limitations, especially when drawing from post hoc analyses based on existing information [36]. Although GDSI showcased significant advancements, challenges inherent to its structure and execution were encountered. In this section, these challenges are delineated, encompassing aspects from data collection and analysis to concerns of interoperability, data quality, governance, data sharing, and privacy. By exploring these areas, insights are provided to optimize future initiatives and fully harness the potential of RWD in the context of global collaborative learning.

Challenges and Solutions in Data Interoperability, Quality, and Governance

Interoperability and handling heterogenous data formats presented significant hurdles for GDSI. To counteract these challenges, a study-specific data dictionary was created. However, more advanced standardization methods such as a common data model [37], including Fast Healthcare Interoperability Resources [38] and Observational Medical Outcomes Partnership [39], could further enhance standardization, making it more generalized and disease-agnostic [40,41]. Building on the necessity for standardization, the significance of data quality has been universally recognized in health care, as also highlighted by various studies [42-45]. In tandem with standardization efforts, GDSI integrated an automated data quality assessment framework into its data acquisition process. However, the adoption of a generalized framework such as [46] can serve as a blueprint for enhancing data quality across various health care contexts, providing a more structured format to ensure reliability and precision. As GDSI confronted challenges related to interoperability and data quality, the initiative also had to navigate the complex landscape of regulatory compliance. Implementing a federated governance model, GDSI effectively addressed the existing needs but simultaneously revealed a gap for a data governance model in health care, namely, the absence of implementations specifically tailored for a federated framework [47]. A more universal data governance model such as the one proposed by Peregrina et al [48] could potentially fill this gap, enhancing both organizational efficiency and the quality of analytical models.

Embracing Federated Model Sharing and Privacy Concerns

Although federated model sharing offers a unique approach to draw insights from patient-level data, it is worth acknowledging that even the impersonal shared statistics inherently encode some information [49]. However, these potential risks are managed under the strict supervision of GDSI, which operates within a rigorously regulated and controlled environment with trusted partners. To further mitigate risks, the data custodians in the federated model sharing underwent a formal assessment of privacy risks after running the script and before sharing the aggregated data with the central platform. This additional layer of scrutiny ensured that any potential privacy concerns were addressed prior to data dissemination. Potential risks and their mitigation strategies were transparently communicated to all data providers via a clear analysis plan, thereby striking a robust balance between efficient data use and strict privacy and security standards. Although GDSI's federated model sharing has proven successful, it falls short in one crucial area: iterative asynchronous communication. This oversight leads to the introduction of federated learning [50], a methodology wherein a machine learning algorithm extracts knowledge from a variety of locally stored data without the need to transfer raw data enabling deploying sophisticated analysis [51]. Nonetheless, it is vital to recognize the associated risks and challenges. Federated learning or, in general, federated model sharing is not invulnerable to attacks [52,53] or privacy breaches [49].

Considering these risks, it might be necessary to re-evaluate GDSI's current assumptions of trustworthiness, inquisitiveness, and nonantagonistic behavior among all participants for a wider scope of application. Incorporating privacy-preserving algorithms such as differential privacy [54] and homomorphic encryption [55] can bolster security measures, though potentially affecting analytical performance or necessitating extensive computational resources [56]. Despite these challenges, federated learning has shown promise in a range of studies [57-61]. However, most of these analyses were tailored to specific use cases. There remains a need for a more generalized federated learning pipeline that can be applied broadly, rather than being limited to project-specific applications [62].

Recognizing the inherent risks in the federated approach, GDSI took proactive steps to ensure privacy and build trust within the entire pipeline. In response to concerns regarding privacy and tool reliability, GDSI adopted privacy-by-design principles and utilized certified toolboxes that underwent third-party verification. This approach emphasizes the continual need for assessment and evaluation of privacy safeguards.

Enhancing Collaboration: User Engagement in the GDSI Pipeline

As GDSI delved deeper into privacy and security measures, it became evident that an improved user experience was pivotal for the pipeline's success. The intricate nature of the RWD analysis pipeline, coupled with its limited visualization capabilities and an initial oversight in stakeholder inclusion, gave rise to a black box perception. Recognizing the urgent need for better communication and more user-friendly tools, GDSI instituted a dedicated task force. This team took charge from the study's inception to the formulation of evidence-based guidelines, guaranteeing that every stage aligned with the multifaceted needs of all stakeholders. By doing so, GDSI not only fostered trust and collaboration but also strongly resonated with the project's overarching principles of engagement and transparency.

The deployment of GDSI's user-centric interactive web application, complemented by detailed documentation and illustrative visuals, helped demystify the pipeline's complexity. By offering accessible and user-friendly tools, this approach fostered a more nuanced stakeholder engagement, bridging the divide between intricate operations and approachability. The effectiveness of visualization in health care is supported by various studies [63-66]. Tools such as Jaspersoft [67], Tableau [68], Looker [69], Domo [70], Tibco Spotfire [71], and Power BI [71] offer a business-level data analytics platform, underscoring the significance of converting intricate data sets into comprehensible visuals.

Pragmatism in GDSI: Balancing Innovation and Adaptation

In the conceptualization and development of GDSI, striking a balance between advanced innovation and practical inclusivity was paramount. This principle was clearly manifested in the design of the data acquisition architecture. Typically, health care frameworks gravitate toward a federated or centralized model. Yet, GDSI embraced a hybrid strategy, seeking to cater

to a broad spectrum of users and registries. This novel approach marked a significant departure from the norm, merging technological advancement with operational flexibility.

However, with innovation comes challenges. Although GDSI's analysis pipeline presents a viable technical solution for collaborative health care learning, it also grappled with broader societal challenges. One salient example was the containerization strategy. Initially promising, it met resistance from certain federated model–sharing registries because of their internal policies. Such challenges underscore the ever-present demand for adaptability amid rapid technological shifts. However, GDSI responded proactively, making the source code available and bolstering it with a comprehensive manual and robust support. Such measures exemplify GDSI's commitment to reconciling groundbreaking advancements with real-world constraints.

This commitment extended beyond technical challenges. The global reach of GDSI emphasized the importance of resource efficiency, especially in regions with limited internet connectivity. In striving for a global impact, GDSI reiterated its pledge to balance technological progress with practical considerations across diverse geographies. In light of these experiences, one thing becomes clear for the success of initiatives like GDSI: continuous education, proactive stakeholder engagement, and evidence-based demonstrations in controlled environments are not just beneficial, but they are essential.

GDSI as a Blueprint for Data-Sharing Initiatives in Biomedical Research

GDSI emerged as a pragmatic blueprint for interdisciplinary biomedical research. The meticulous planning and systematic execution of the initiative showcased how strategic processes can serve as foundational guides for upcoming biomedical consortia. The open-source resources GDSI provides [23,25,29,30,32,34,35,72-74] can be directly leveraged and adjusted after thorough assessment and evaluation. These resources bifurcate into 2 main categories: disease-agnostic and disease-specific components.

Within the context of disease-agnostic components, the architecture of GDSI's end-to-end data analysis pipeline stands out, highlighting its modularity and adaptability. This pipeline's design facilitates significant customization, catering to various data acquisition streams. The hybrid nature of the data acquisition module allows initiatives to choose one or a combination of data collection methods based on their distinct needs and policies. Additionally, GDSI's data integration framework plays a crucial role in amalgamating these diverse data sources into a unified and comprehensive data set. Together, these components offer a versatile foundation that other biomedical initiatives can adapt and leverage according to their specific requirements.

Turning to disease-specific components, aspects like the data dictionary and data quality assessments were designed primarily for the research question centered around MS and COVID-19. Even though these components are specialized, they act as guiding principles for other research ventures. The data dictionary, augmented by its metadata, provides a robust foundation for the next phases of the pipeline. It offers a detailed account of acquisition variables and sets clear data quality criteria. A significant point to note is that the data dictionary aids in determining the rules for data quality assessments, presenting a methodical approach to data validation. This thorough approach emphasizes the importance of precise planning and specificity when delving into disease-focused research questions, setting an example for other initiatives to follow.

To conclude, the flexibility and adaptability inherent in GDSI's comprehensive data analysis pipeline coupled with its disease-specific components meld to present a versatile tool for crafting sturdy data architectures across a spectrum of biomedical research landscapes. Those seeking a deeper understanding and guidance on harnessing and replicating GDSI's capabilities can refer to Multimedia Appendix 3, which offers a detailed roadmap based on GDSI's experiences and insights, a flowchart tracing the initiative from its inception to its research question resolutions, and guidance on replicating GDSI's federated model sharing infrastructure. A graphical abstract delineating the high-level architecture of this study is presented in Multimedia Appendix 4, which provides additional insights regarding the architectural framework.

Conclusion

GDSI had substantial impact that extended beyond its initial focus on COVID-19 and MS. It contributed to numerous scientific publications and played a pivotal role in shaping global guidelines for the community with MS [7,9,16-19]. This underscores the vast potential of data-driven collaborative efforts to yield improved health care outcomes. A cornerstone of GDSI's success was its RWD analysis pipeline. Crafted to navigate technical, epidemiological, and sociological challenges, this pipeline facilitated the seamless integration of varied data streams into a single data set. This cohesive strategy enabled large-scale collaborative research and offered the flexibility to accommodate the diverse policies, regulations, and needs of various data providers. Serving as a practical blueprint, GDSI addressed not only current health care challenges but also laid the groundwork for future initiatives. Its hybrid approach to data acquisition and analysis provided a scalable framework applicable to other health care sectors. In doing so, GDSI stands as a compelling example of how data sharing and collaborative learning can meaningfully advance health care research, going beyond the specific challenges of MS and COVID-19.

Acknowledgments

The author(s) have disclosed that they received financial support for the research, authorship, or publication of this paper from the following sources: the operational costs associated with this study were funded by the Multiple Sclerosis International Federation and the Multiple Sclerosis Data Alliance (MSDA) operating under the European Charcot Foundation. The MSDA is a global not-for-profit multistakeholder collaboration acting under the umbrella of the European Charcot Foundation, financially supported by a combination of industry partners, including Novartis, Merck, Biogen, Janssen, Bristol-Myers Squibb, and Roche. Additionally, this work was supported by the Flemish government through the Onderzoeksprogramma Artificiële Intelligentie Vlaanderen program and the Research Foundation Flanders for ELIXIR Belgium. QMENTA provided the central platform, while Amazon supplied the computational resources utilized in this work. The statistical analysis was conducted at the Clinical Outcomes Research Unit, The University of Melbourne, with support from National Health and Medical Research Council (1129189 and 1140766). The authors wish to extend their sincere appreciation to Nikola Lazovski for his invaluable guidance and collaboration throughout the global data sharing initiative project, especially concerning the central platform. They are also profoundly grateful to Dr Ilse Vermeulen for her unwavering support and encouragement throughout the various stages of drafting and conceptualizing the manuscript.

Authors' Contributions

AP played a major role in acquiring the data, designing the federated infrastructures, and drafting and revising the manuscript for content. LG and TP contributed to manuscript drafting and revision for content. EDB contributed to manuscript drafting and revision, played a major role in acquiring data, and contributed to the study concept or design as well as the analysis or interpretation of data. YM and LMP provided overall supervision and coordination of the study and contributed to manuscript drafting and revision for content, including writing for content.

Conflicts of Interest

AP and YM are funded by VLAIO (Flanders Innovation and Entrepreneurship) PM: Augmenting Therapeutic Effectiveness through Novel Analytics (HBC.2019.2528) and Research Council Katholieke Universiteit Leuven: Symbiosis 4 (C14/22/125) and Symbiosis 3 (C14/18/092; CELSA - Active Learning; CELSA/21/019). AP and YM are affiliated to Leuven.AI and received funding from the Flemish government (Artificial Intelligence Research Program, Research Foundation Flanders [FWO] Strategic Basic [SB] Research; S003422N). EDB was funded by an FWO-SB grant. LMP is the chair of MSDA, which received income from a range of corporate sponsors, including Biogen, Bristol Myers Squibb, Janssen Pharmaceuticals, Merck, Novartis, and Roche. LG is funded by the Flemish government under the Onderzoeksprogramma Artificiele Intelligentie Vlaanderen. TP is funded by the Flemish government under the Bijzonder Onderzoeksfonds special research fund BOF22OWB01.

Multimedia Appendix 1
Analysis of the data dictionary employed within the global data sharing initiative.
[DOCX File , 20 KB - medinform_v11i1e48030_app1.docx]

Multimedia Appendix 2
The 3 iterative pipelines developed: federated pipeline COV1.0, COV2.0, and COV2.1.
[DOCX File , 139 KB - medinform_v11i1e48030_app2.docx]

Multimedia Appendix 3
Roadmap of the global data sharing initiative.
[DOCX File , 427 KB - medinform_v11i1e48030_app3.docx]

Multimedia Appendix 4
Graphical abstract.
[PNG File , 3416 KB - medinform_v11i1e48030_app4.png]

References

1. Mitani AA, Haneuse S. Small data challenges of studying rare diseases. JAMA Netw Open 2020 Mar 02;3(3):e201965 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.1965] [Medline: 32202640]
2. Walton C, King R, Rechtman L, Kaye W, Leray E, Marrie RA, et al. Rising prevalence of multiple sclerosis worldwide: Insights from the Atlas of MS, third edition. Mult Scler 2020 Dec;26(14):1816-1821 [FREE Full text] [doi: 10.1177/1352458520970841] [Medline: 33174475]
3. Real-world evidence. US Food and Drug Administration. URL: https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence [accessed 2023-03-30]

4. Burcu M, Dreyer NA, Franklin JM, Blum MD, Critchlow CW, Perfetto EM, et al. Real-world evidence to support regulatory decision-making for medicines: considerations for external control arms. Pharmacoepidemiol Drug Saf 2020 Oct;29(10):1228-1235 [FREE Full text] [doi: 10.1002/pds.4975] [Medline: 32162381]
5. Rudrapatna VA, Butte AJ. Opportunities and challenges in using real-world data for health care. J Clin Invest 2020 Feb 03;130(2):565-574 [FREE Full text] [doi: 10.1172/JCI129197] [Medline: 32011317]
6. Hiramatsu K, Barrett A, Miyata Y, PhRMA Japan Medical Affairs Committee Working Group 1. Current status, challenges, and future perspectives of real-world data and real-world evidence in Japan. Drugs Real World Outcomes 2021 Dec;8(4):459-480 [FREE Full text] [doi: 10.1007/s40801-021-00266-3] [Medline: 34148219]
7. Simpson-Yap S, De Brouwer E, Kalincik T, Rijke N, Hillert JA, Walton C, et al. Associations of disease-modifying therapies with COVID-19 severity in multiple sclerosis. Neurology 2021 Nov 09;97(19):e1870-e1885 [FREE Full text] [doi: 10.1212/WNL.0000000000012753] [Medline: 34610987]
8. Katkade VB, Sanders KN, Zou KH. Real world data: an opportunity to supplement existing evidence for the use of long-established medicines in health care decision making. JMDH 2018 Jul;Volume 11:295-304. [doi: 10.2147/jmdh.s160029]
9. Peeters LM, Parciak T, Walton C, Geys L, Moreau Y, De Brouwer E, et al. COVID-19 in people with multiple sclerosis: a global data sharing initiative. Mult Scler 2020 Sep;26(10):1157-1162 [FREE Full text] [doi: 10.1177/1352458520941485] [Medline: 32662757]
10. Antoniou V, Vassilakis E, Hatzaki M. Is crowdsourcing a reliable method for mass data acquisition? The case of COVID-19 spread in Greece during spring 2020. ISPRS Int J Geo Inf 2020 Oct 14;9(10):605. [doi: 10.3390/ijgi9100605]
11. Yu C, Chang S, Chang T, Wu JL, Lin Y, Chien H, et al. A COVID-19 pandemic artificial intelligence-based system with deep learning forecasting and automatic statistical data acquisition: development and implementation study. J Med Internet Res 2021 May 20;23(5):e27806 [FREE Full text] [doi: 10.2196/27806] [Medline: 33900932]
12. Khalid S, Yang C, Blacketer C, Duarte-Salles T, Fernández-Bertolín S, Kim C, et al. A standardized analytics pipeline for reliable and rapid development and validation of prediction models using observational health data. Comput Methods Programs Biomed 2021 Nov;211:106394 [FREE Full text] [doi: 10.1016/j.cmpb.2021.106394] [Medline: 34560604]
13. Nishimwe A, Ruranga C, Musanabaganwa C, Mugeni R, Semakula M, Nzabanita J, et al. Leveraging artificial intelligence and data science techniques in harmonizing, sharing, accessing and analyzing SARS-COV-2/COVID-19 data in Rwanda (LAISDAR Project): study design and rationale. BMC Med Inform Decis Mak 2022 Aug 12;22(1):214 [FREE Full text] [doi: 10.1186/s12911-022-01965-9] [Medline: 35962355]
14. Junior EPP, Normando P, Flores-Ortiz R, Afzal MU, Jamil MA, Bertolin SF, et al. Integrating real-world data from Brazil and Pakistan into the OMOP common data model and standardized health analytics framework to characterize COVID-19 in the Global South. J Am Med Inform Assoc 2023 Mar 16;30(4):643-655 [FREE Full text] [doi: 10.1093/jamia/ocac180] [Medline: 36264262]
15. Peeters LM, Parciak T, Kalra D, Moreau Y, Kasilingam E, van Galen P, et al. Multiple Sclerosis Data Alliance - A global multi-stakeholder collaboration to scale-up real world data research. Mult Scler Relat Disord 2021 Jan;47:102634 [FREE Full text] [doi: 10.1016/j.msard.2020.102634] [Medline: 33278741]
16. Simpson-Yap S, Brouwer ED, Kalincik T, et al. Associations of DMT therapies with COVID-19 severity in multiple sclerosis? Int J Epidemiol 2021 Sep 02:51. [doi: 10.1093/ije/dyab168.604]
17. Simpson-Yap S, Pirmani A, Kalincik T, De Brouwer E, Geys L, Parciak T, et al. Updated results of the COVID-19 in MS global data sharing initiative. Neurol Neuroimmunol Neuroinflamm 2022 Aug 29;9(6):e200021. [doi: 10.1212/nxi.0000000000200021]
18. An update from the MS global data sharing initiative - thank you!. MS International Federation. URL: https://www.msif.org/news/2022/04/06/an-update-from-the-ms-global-data-sharing-initiative-thank-you/ [accessed 2023-03-30]
19. Simpson-Yap S, Pirmani A, De Brouwer E, Peeters LM, Geys L, Parciak T, et al. Severity of COVID19 infection among patients with multiple sclerosis treated with interferon-β. Mult Scler Relat Disord 2022 Oct;66:104072 [FREE Full text] [doi: 10.1016/j.msard.2022.104072] [Medline: 35917745]
20. McCabe A, Nic An Fhailí S, O'Sullivan R, Brenner M, Gannon B, Ryan J, et al. Development and validation of a data dictionary for a feasibility analysis of emergency department key performance indicators. Int J Med Inform 2019 Jun;126:59-64. [doi: 10.1016/j.ijmedinf.2019.01.015] [Medline: 31029264]
21. Lin S, Morrison LJ, Brooks SC. Development of a data dictionary for the Strategies for Post Arrest Resuscitation Care (SPARC) network for post cardiac arrest research. Resuscitation 2011 Apr;82(4):419-422. [doi: 10.1016/j.resuscitation.2010.12.006] [Medline: 21276647]
22. Moss E. The national health data dictionary. Health Inf Manag 1994 Mar;24(1):26-29. [doi: 10.1177/183335839402400112] [Medline: 10141009]
23. COVID19-GDSI/data dictionary. GitHub. URL: https://github.com/MS-DATA-ALLIANCE/COVID19-GDSI/blob/main/Data%20Dictionary.pdf [accessed 2023-04-09]
24. Cook JA, Collins GS. The rise of big clinical databases. Br J Surg 2015 Jan;102(2):e93-e101. [doi: 10.1002/bjs.9723] [Medline: 25627139]
25. COVID19-GDSI/Buckets for federated registries. GitHub. URL: https://github.com/MS-DATA-ALLIANCE/COVID19-GDSI/blob/main/Buckets%20for%20federated%20registries.pdf [accessed 2023-04-11]

26. Anderson C. Docker [software engineering]. IEEE Softw 2015 May;32(3):102-1c3. [doi: 10.1109/ms.2015.62]
27. Alpine Linux. URL: https://www.alpinelinux.org/ [accessed 2023-07-22]
28. Demo presentation of the federated pipeline. MSDA Federated Pipeline COV 2.1 YouTube page. URL: https://www.youtube.com/watch?v=d-QuCNDbHKc&ab_channel=AshkanPirmani [accessed 2023-03-31]
29. MS-DATA-ALLIANCE/COVID19-GDSI. GitHub. URL: https://github.com/MS-DATA-ALLIANCE/COVID19-GDSI/tree/main [accessed 2023-04-09]
30. MS-DATA-ALLIANCE/COVID19-GDSI2021: MSDA toolkit for federated registries participating in the COVID-19 in MS Global Data Sharing initiative (GDSI). GitHub. URL: https://github.com/MS-DATA-ALLIANCE/COVID19-GDSI2021 [accessed 2023-03-31]
31. msdaalliance/covid19gdsi-ui. Docker Hub. URL: https://hub.docker.com/r/msdaalliance/covid19gdsi-ui [accessed 2023-03-31]
32. MS-DATA-ALLIANCE/COVID19-GDSI2021-UI: MSDA toolkit for federated registries participating in the COVID-19 in MS Global Data Sharing initiative (GDSI). GitHub. URL: https://github.com/MS-DATA-ALLIANCE/COVID19-GDSI2021-UI [accessed 2023-04-11]
33. Snyk. URL: https://snyk.io/ [accessed 2023-03-31]
34. MS-DATA-ALLIANCE/COVID19-GDSI. GitHub. URL: https://github.com/MS-DATA-ALLIANCE/COVID19-GDSI/blob/main/Data%20quality%20assessment%20and%20enhancement%20pipeline.pdf [accessed 2023-04-09]
35. Patient-level dataset to study the effect of COVID-19 in people with multiple sclerosis v1. PhysionNet. URL: https://physionet.org/content/patient-level-data-covid-ms/1.0.0/ [accessed 2023-08-13]
36. Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. N Engl J Med 2000 Jun 22;342(25):1878-1886. [doi: 10.1056/nejm200006223422506]
37. Voss EA, Makadia R, Matcho A, Ma Q, Knoll C, Schuemie M, et al. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. J Am Med Inform Assoc 2015 May;22(3):553-564 [FREE Full text] [doi: 10.1093/jamia/ocu023] [Medline: 25670757]
38. Pfaff ER, Champion J, Bradford RL, Clark M, Xu H, Fecho K, et al. Fast healthcare interoperability resources (FHIR) as a meta model to integrate common data models: development of a tool and quantitative validation study. JMIR Med Inform 2019 Oct 16;7(4):e15199 [FREE Full text] [doi: 10.2196/15199] [Medline: 31621639]
39. Ahmadi N, Peng Y, Wolfien M, Zoch M, Sedlmayr M. OMOP CDM can facilitate data-driven studies for cancer prediction: a systematic review. Int J Mol Sci 2022 Oct 05;23(19):11834 [FREE Full text] [doi: 10.3390/ijms231911834] [Medline: 36233137]
40. Yu Y, Zong N, Wen A, Liu S, Stone DJ, Knaack D, et al. Developing an ETL tool for converting the PCORnet CDM into the OMOP CDM to facilitate the COVID-19 data integration. J Biomed Inform 2022 Mar;127:104002 [FREE Full text] [doi: 10.1016/j.jbi.2022.104002] [Medline: 35077901]
41. Makadia R, Ryan PB. Transforming the premier perspective hospital database into the observational medical outcomes partnership (OMOP) common data model. EGEMS (Wash DC) 2014;2(1):1110 [FREE Full text] [doi: 10.13063/2327-9214.1110] [Medline: 25848597]
42. Chao-Gan Y, Yu-Feng Z. DPARSF: A MATLAB toolbox for "pipeline" data analysis of resting-state fMRI. Front Syst Neurosci 2010;4:13 [FREE Full text] [doi: 10.3389/fnsys.2010.00013] [Medline: 20577591]
43. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, et al. From FastQ data to high confidence variant calls: the genome analysis toolkit best practices pipeline. Curr Protoc Bioinformatics 2013;43(1110):11.10.1-11.10.33 [FREE Full text] [doi: 10.1002/0471250953.bi1110s43] [Medline: 25431634]
44. Uritskiy GV, DiRuggiero J, Taylor J. MetaWRAP-a flexible pipeline for genome-resolved metagenomic data analysis. Microbiome 2018 Sep 15;6(1):158 [FREE Full text] [doi: 10.1186/s40168-018-0541-1] [Medline: 30219103]
45. Chen H, Lau MC, Wong MT, Newell EW, Poidinger M, Chen J. Cytofkit: A bioconductor package for an integrated mass cytometry data analysis pipeline. PLoS Comput Biol 2016 Sep;12(9):e1005112 [FREE Full text] [doi: 10.1371/journal.pcbi.1005112] [Medline: 27662185]
46. Lee K, Weiskopf N, Pathak J. A framework for data quality assessment in clinical research datasets. AMIA Annu Symp Proc 2017;2017:1080-1089 [FREE Full text] [Medline: 29854176]
47. Perez JA, Bellot GO, Zirpins C. Data governance for federated machine learning in secure web-based systems. In: Minutes of the Predoctoral Research Conference in Computer Engineering: Proceedings of the Doctoral Consortium in Computer Science. Spain: Universidad de la Rioja; 2021:36-39.
48. Peregrina JA, Ortiz G, Zirpins C. Towards data governance for federated machine learning. In: Advances in Service-Oriented and Cloud Computing. Switzerland: Springer Cham; 2022.
49. Nasirigerdeh R, Torkzadehmahani R, Baumbach J, et al. On the privacy of federated pipelines. 2021 Presented at: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval; July 11-15; Canada. [doi: 10.1145/3404835.3462996]
50. McMahan B, Moore E, Ramage D, Hampson S, et al. Communication-efficient learning of deep networks from decentralized data. arXiv Preprint posted online on February 17, 2016. [doi: 10.48550/arXiv.1602.05629]
51. Rieke N, Hancox J, Li W, Milletarì F, Roth HR, Albarqouni S, et al. The future of digital health with federated learning. NPJ Digit Med 2020;3:119 [FREE Full text] [doi: 10.1038/s41746-020-00323-1] [Medline: 33015372]

52. Bagdasaryan E, Veit A, Hua Y, et al. How to backdoor federated learning. arXiv Preprint posted online on July 2, 2018. [doi: 10.48550/arXiv.1807.00459]
53. Tolpegin V, Truex S, Gursoy ME, Liu L. Data poisoning attacks against federated learning systems. arXiv Preprint posted online on August 11, 2020. [doi: 10.1007/978-3-030-58951-6_24]
54. Cynthia D. Differential privacy. In: Lecture Notes in Computer Science. Switzerland: Springer Nature; 2006:1-12.
55. Wood A, Najarian K, Kahrobaei D. Homomorphic encryption for machine learning in medicine and bioinformatics. ACM Comput Surv 2020 Aug 25;53(4):1-35. [doi: 10.1145/3394658]
56. Raisaro JL, Choi G, Pradervand S, Colsenet R, Jacquemont N, Rosat N, et al. Protecting privacy and security of genomic data in i2b2 with homomorphic encryption and differential privacy. IEEE/ACM Trans. Comput. Biol. and Bioinf 2018:1-1. [doi: 10.1109/tcbb.2018.2854782]
57. Deist TM, Dankers FJWM, Ojha P, Scott Marshall M, Janssen T, Faivre-Finn C, et al. Distributed learning on 20,000+ lung cancer patients - The personal health train. Radiother Oncol 2020 Mar;144:189-200 [FREE Full text] [doi: 10.1016/j.radonc.2019.11.019] [Medline: 31911366]
58. Huang L, Shea AL, Qian H, Masurkar A, Deng H, Liu D. Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records. J Biomed Inform 2019 Nov;99:103291 [FREE Full text] [doi: 10.1016/j.jbi.2019.103291] [Medline: 31560949]
59. Jochems A, Deist TM, El Naqa I, Kessler M, Mayo C, Reeves J, et al. Developing and validating a survival prediction model for NSCLC patients through distributed learning across 3 countries. Int J Radiat Oncol Biol Phys 2017 Oct 01;99(2):344-352 [FREE Full text] [doi: 10.1016/j.ijrobp.2017.04.021] [Medline: 28871984]
60. Li J, Tian Y, Zhu Y, Zhou T, Li J, Ding K, et al. A multicenter random forest model for effective prognosis prediction in collaborative clinical research network. Artif Intell Med 2020 Mar;103:101814. [doi: 10.1016/j.artmed.2020.101814] [Medline: 32143809]
61. Li X, Gu Y, Dvornek N, Staib LH, Ventola P, Duncan JS. Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results. Med Image Anal 2020 Oct;65:101765 [FREE Full text] [doi: 10.1016/j.media.2020.101765] [Medline: 32679533]
62. Pirmani A, De Brouwer E, Moreau Y, Peeters LM. Federated learning for everyone (FL4E). Elixir All Hands Meeting 2023:1. [doi: 10.7490/F1000RESEARCH.1119405.1]
63. Gotz D, Borland D. Data-driven healthcare: challenges and opportunities for interactive visualization. IEEE Comput Grap Appl 2016 May;36(3):90-96. [doi: 10.1109/mcg.2016.59]
64. Stadler JG, Donlon K, Siewert JD, Franken T, Lewis NE. Improving the efficiency and ease of healthcare analysis through use of data visualization dashboards. Big Data 2016 Jun;4(2):129-135. [doi: 10.1089/big.2015.0059] [Medline: 27441717]
65. Menon A, Aishwarya MS, Joykutty AM, Av AY. Data visualization and predictive analysis for smart healthcare: tool for a hospital. 2021 Presented at: 2021 IEEE Region 10 Symposium (TENSYMP); October 4, 2021; Jeju, Republic of Korea p. 1-8. [doi: 10.1109/TENSYMP52854.2021.9550822]
66. Battineni G, Mittal M, Jain S. Data visualization in the transformation of healthcare industries. In: Advanced Prognostic Predictive Modelling in Healthcare Data Analytics. Singapore: Springer; 2021:1-23.
67. Reporting and embedded business intelligence software. Jaspersoft. URL: https://www.jaspersoft.com/ [accessed 2023-07-31]
68. Business intelligence and analytics software. Tableau. URL: https://www.tableau.com/ [accessed 2023-07-31]
69. Looker business intelligence platform embedded analytics. Google Cloud. URL: https://cloud.google.com/looker [accessed 2023-08-11]
70. Discover the Domo data experience platform. Domo. URL: https://www.domo.com/ [accessed 2023-08-11]
71. Data visualization. Microsoft Power BI. URL: https://powerbi.microsoft.com/en-us/ [accessed 2023-08-11]
72. jupyter/datascience-notebook. Docker Hub. URL: https://hub.docker.com/r/jupyter/datascience-notebook [accessed 2023-03-31]
73. .NET. Microsoft. URL: https://dotnet.microsoft.com/en-us/ [accessed 2023-07-31]
74. The cron schedule expression editor. Crontab. URL: https://crontab.guru/ [accessed 2023-07-31]

Abbreviations

aOR: adjusted odds ratio
DMT: disease-modifying therapy
GDSI: global data sharing initiative
MS: multiple sclerosis
RWD: real-world data

Edited by C Lovis; submitted 20.04.23; peer-reviewed by X Liu, D Blumenthal; comments to author 05.07.23; revised version received 25.08.23; accepted 30.09.23; published 09.11.23.

Please cite as:
Pirmani A, De Brouwer E, Geys L, Parciak T, Moreau Y, Peeters LM
The Journey of Data Within a Global Data Sharing Initiative: A Federated 3-Layer Data Analysis Pipeline to Scale Up Multiple Sclerosis Research
JMIR Med Inform 2023;11:e48030
URL: https://medinform.jmir.org/2023/1/e48030
doi:10.2196/48030
PMID:37943585

Computer-Supported Collaborative Design of Standardized Clinical Cases: Algorithm Development and Validation

Sergio Guinez-Molinos[1,*], MSc, PhD; Félix Buendía-García[2,*], PhD; José-Luis Sierra-Rodríguez[3,*], PhD; Joaquín Gayoso-Cabada[4,*], PhD; Jaime González-Díaz[1,*], MSc

1

2

3

4

*all authors contributed equally

Corresponding Author:
Sergio Guinez-Molinos, MSc, PhD

Abstract

Background: The creation of computer-supported collaborative clinical cases is an area of educational research that has been widely studied. However, the reuse of cases and their sharing with other platforms is a problem, as it encapsulates knowledge in isolated platforms without interoperability. This paper proposed a workflow ecosystem for the collaborative design and distribution of clinical cases through web-based computing platforms that (1) allow medical students to create clinical cases collaboratively in a dedicated environment; (2) make it possible to export these clinical cases in terms of the Health Level 7 (HL7) Fast Healthcare Interoperability Resources (FHIR) interoperability standard; (3) provide support to transform imported cases into learning object repositories; and (4) use e-learning standards (eg, Instructional Management Systems Content Packaging [IMS-CP] or Sharable Content Object Reference Model [SCORM]) to incorporate this content into widely-used learning management systems (LMSs), letting medical students democratize a valuable knowledge that would otherwise be confined within proprietary platforms.

Objective: This study aimed to demonstrate the feasibility of developing a workflow ecosystem based on IT platforms to enable the collaborative creation, export, and deployment of clinical cases.

Methods: The ecosystem infrastructure for computer-supported collaborative design of standardized clinical cases consists of three platforms: (1) Mosaico, a platform used in the design of clinical cases; (2) Clavy, a tool for the flexible management of learning object repositories, which is used to orchestrate the transformation and processing of these clinical cases; and (3) Moodle, an LMS that is geared toward publishing the processed clinical cases and delivering their course deployment stages in IMS-CP or SCORM format. The generation of cases in Mosaico is exported in the HL7 FHIR interoperability standard to Clavy, which is then responsible for creating and deploying a learning object in Moodle.

Results: The main result was an interoperable ecosystem that demonstrates the feasibility of automating the stages of collaborative clinical case creation, export through HL7 FHIR standards, and deployment in an LMS. This ecosystem enables the generation of IMS-CPs associated with the original Mosaico clinical cases that can be deployed in conventional third-party LMSs, thus allowing the democratization and sharing of clinical cases to different platforms in standard and interoperable formats.

Conclusions: In this paper, we proposed, implemented, and demonstrated the feasibility of developing a standards-based workflow that interoperates multiple platforms with heterogeneous technologies to create, transform, and deploy clinical cases on the web. This achieves the objective of transforming the created cases into a platform for web-based deployment in an LMS.

(JMIR Med Inform 2023;11:e45315) doi:10.2196/45315

KEYWORDS

collaborative learning; interoperability; case-based learning; standards; clinical cases; collaborative clinical cases

Introduction

In March 2020, the World Health Organization (WHO) declared COVID-19 a pandemic [1], directly affecting traditional teaching in medical schools worldwide [2,3]. As a result, at the beginning of April 2020, classroom training activities were suspended at all universities, and students' clinical training was disrupted by the collapse of public and private hospitals [4]. This situation has prompted new teaching strategies for medical students, who could not attend traditional clinical rotations, representing one of the main challenges in this educational context [5]. In this sense, COVID-19 triggered alternative IT-based medical education methodologies, with synchronous and asynchronous mechanisms to be deployed in web-based learning environments

[6]. Afterward, web-based medical education's need for content production has quickly become apparent.

Computer-supported, didactic medical content production in web-based environments demands innovative platforms that enable collaboration among medical students and instructors. Mosaico is an example of a web platform [7] that allows users to design, perform, and assess collaborative clinical scenarios for medical students. Similar platforms have been developed to foster collaborative learning for medical education. For example, Osmosis [8] is a web and mobile learning platform that provides access to multiple clinical questions and explanations for medical student self-assessments. Wikis have also been used as collaborative platforms to build and curate didactic content in graduate medical education [9]. Vicente et al [10] created a platform to enable collaborative learning in the context of the One Health initiative, which aimed to elaborate a corpus of graduate courses or modules in multiple health care disciplines. In addition, Alicanto [11] is a web-based collaboration platform for health care professionals that helps them to share educational resources or discuss clinical cases.

The main advantage of these platforms is to support the learning of groups of students who participate in collaborative activities oriented to generating clinical cases, supporting a didactic approach based on the collaborative design of clinical cases, whose educational value has been widely demonstrated. However, its main shortcoming is that the contents produced are usually oriented to self-consumption within the tools. This restriction limits the reuse of content by not having the ability to export to other platforms, such as conventional learning management systems (LMSs) [12].

Therefore, considering the multiple challenges imposed by the COVID-19 pandemic and the adequacy of practical teaching, it becomes evident that support technology for the teaching-learning processes must allow the export and sharing of the material generated. Thus, we must expand the frontiers of knowledge and open up the consumption of clinical cases created by tools at all levels, starting with LMS platforms. Otherwise, content reuse in contexts other than production tools will be seriously hampered. This paper aimed to address this problem.

We proposed using Mosaico as a platform for the computer-supported collaborative design of educational clinical cases through collaborative clinical simulation [13,14]. To extend and democratize the content generated, Mosaico was extended with an interoperability add-on implemented with an internationally accepted standard, Health Level 7 (HL7) Fast Healthcare Interoperability Resources (FHIR) [15], which allows the designed clinical cases to be exported to a wide variety of third-party platforms. In addition, it allowed us to connect Mosaico with Clavy [16], a tool for managing learning object repositories that enables the transformation into a wide variety of formats and has demonstrated its applicability to the generation of medical learning content from preexisting medical collections. Using Clavy's transformation capabilities, clinical cases retrieved from Mosaico and rendered in HL7 FHIR can be transformed into standardized learning content formats that can be deployed in web-based LMSs such as Moodle or Blackboard, which ultimately enabled us to meet the goal of exporting collaboratively designed Mosaico clinical cases to widely used conventional web-based learning platforms. In this sense, we linked the generation of content with its deployment, sharing it with e-learning environments and extending its application and use by medical students, who can access and share clinical content in a standardized way.

To demonstrate the feasibility of the developed ecosystem, a clinical case of chest pain that considers the whole workflow (collaborative design, transformation, deployment, and consumption) is presented.

Methods

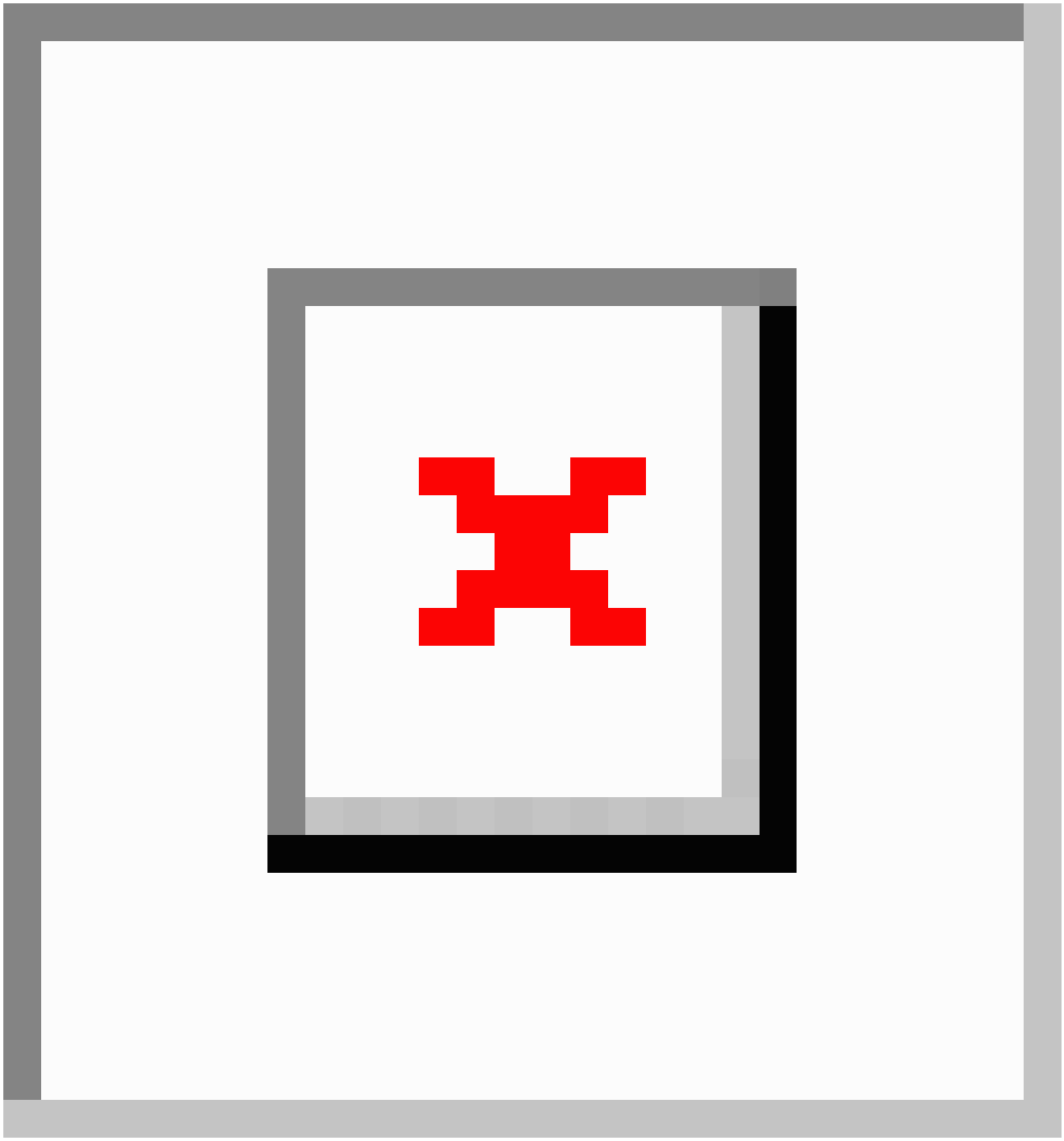
Overview

The collaborative design of clinical cases, created by medical students through a computer-based platform, supports integrating what has been learned and generating knowledge consumed by other students.

The activity diagram of Figure 1 depicts our approach to case-based learning in clinical undergraduate programs. According to this diagram, the approach comprises the following activities:

- *Collaborative design*: In this activity, while facilitated by a reputed instructor, a community of advanced students is involved in the collaborative production of clinical cases. This activity is carried out until the community reaches a clinical case that the instructor judges to be satisfactory.
- *Transformation*: In this activity, the collaboratively produced clinical cases are transformed by instructors into standardized learning packages. To do this, instructors can carry out different types of transformations, which take a given representation of clinical cases as input and generate refined representations as output. These transformations can be carried out automatically, semiautomatically (with instructors configuring some of the automatic processes involved), or manually (with instructors editing content with appropriate editing tools). The process evolves, in general, through several transformation steps until a suitable representation is reached, which is then directly exported into standardized e-learning packages.
- *Deployment*: In this activity, instructors can publish the e-learning packages resulting from the successive transformations into LMSs that support the e-learning standards used to code these content packages.
- *Consumption*: In this activity, less advanced students can take advantage of the conventional e-learning courses generated based on the clinical cases produced in the initial collaborative environment.

Figure 1. Activity diagram for the collaborative production and exploitation of clinical cases.



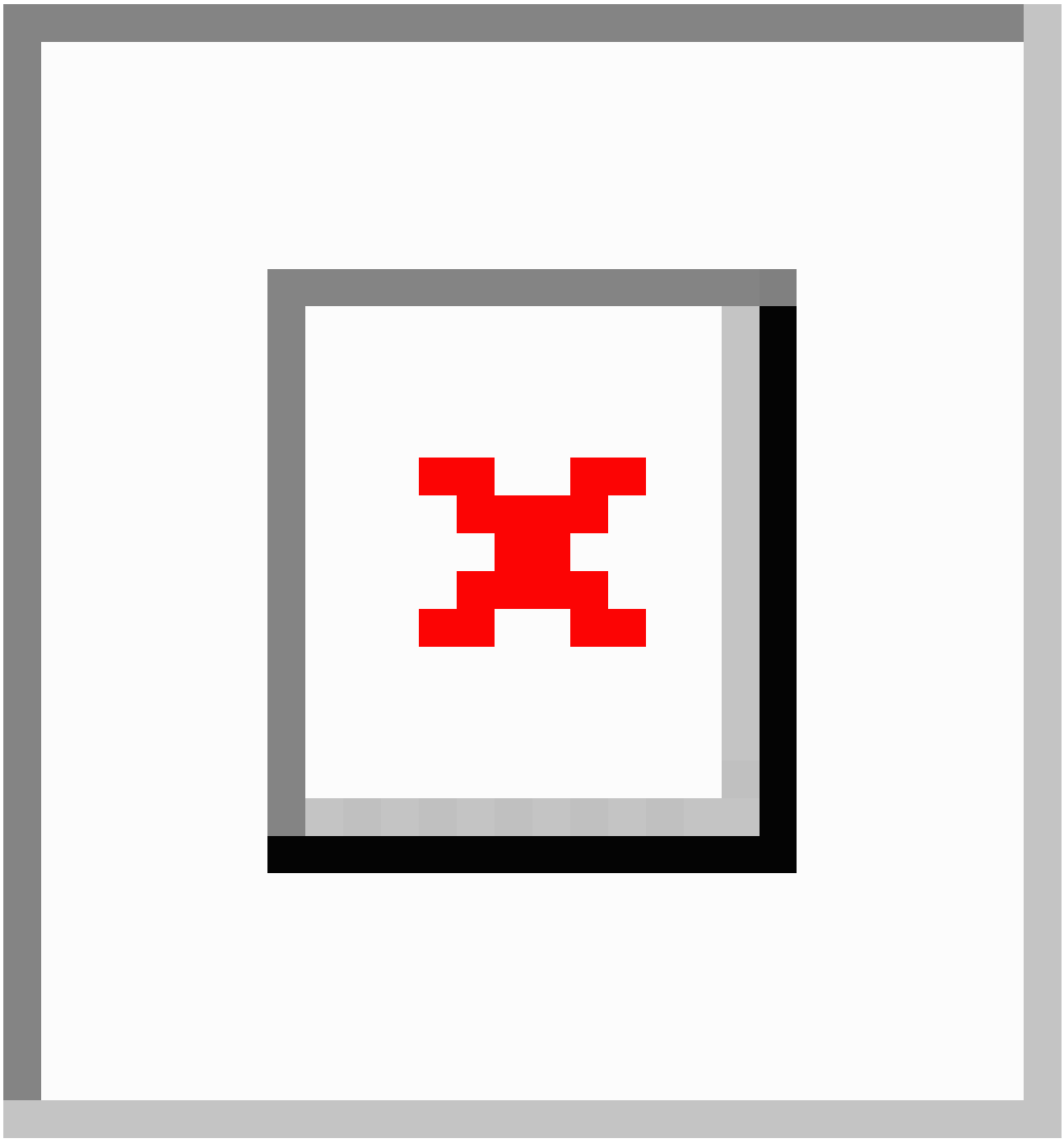
Consequently, learning occurs in two different scenarios:

1. Learning takes place during the collaborative design of clinical cases, which is oriented to the community of advanced learners mediated by the instructor (*Collaborative design* activity).
2. Once clinical cases are transformed by instructors into standardized learning packages, more conventional learning takes place based on content published in a standard LMS (*Consumption* activity).

In this context, we designed and developed a workflow ecosystem based on the collaborative cocreation of clinical cases, supported by technological platforms and global standards that allow for the standardization, export, and distribution of clinical cases for open and democratic consumption. This ecosystem, which is illustrated in Figure 2, integrates the following platforms:

- Mosaico: The platform is used in computer-supported collaborative design of clinical cases [7].
- Clavy: A tool for the flexible management of learning object repositories that is used to orchestrate the transformation stage [16].
- Moodle: The LMS orchestrates the publishing and course consumption stages in the Instructional Management Systems Content Packaging (IMS-CP) or Sharable Content Object Reference Model (SCORM) format [17].

Figure 2. Workflow of the collaborative creation of a clinical case, interoperable with Health Level 7 (HL7) Fast Healthcare Interoperability Resources (FHIR) and packaged as Instructional Management Systems Content Packaging (IMS-CP) or Sharable Content Object Reference Model (SCORM) standard learning objects to be distributed in any learning management system (LMS) that reads these standards.

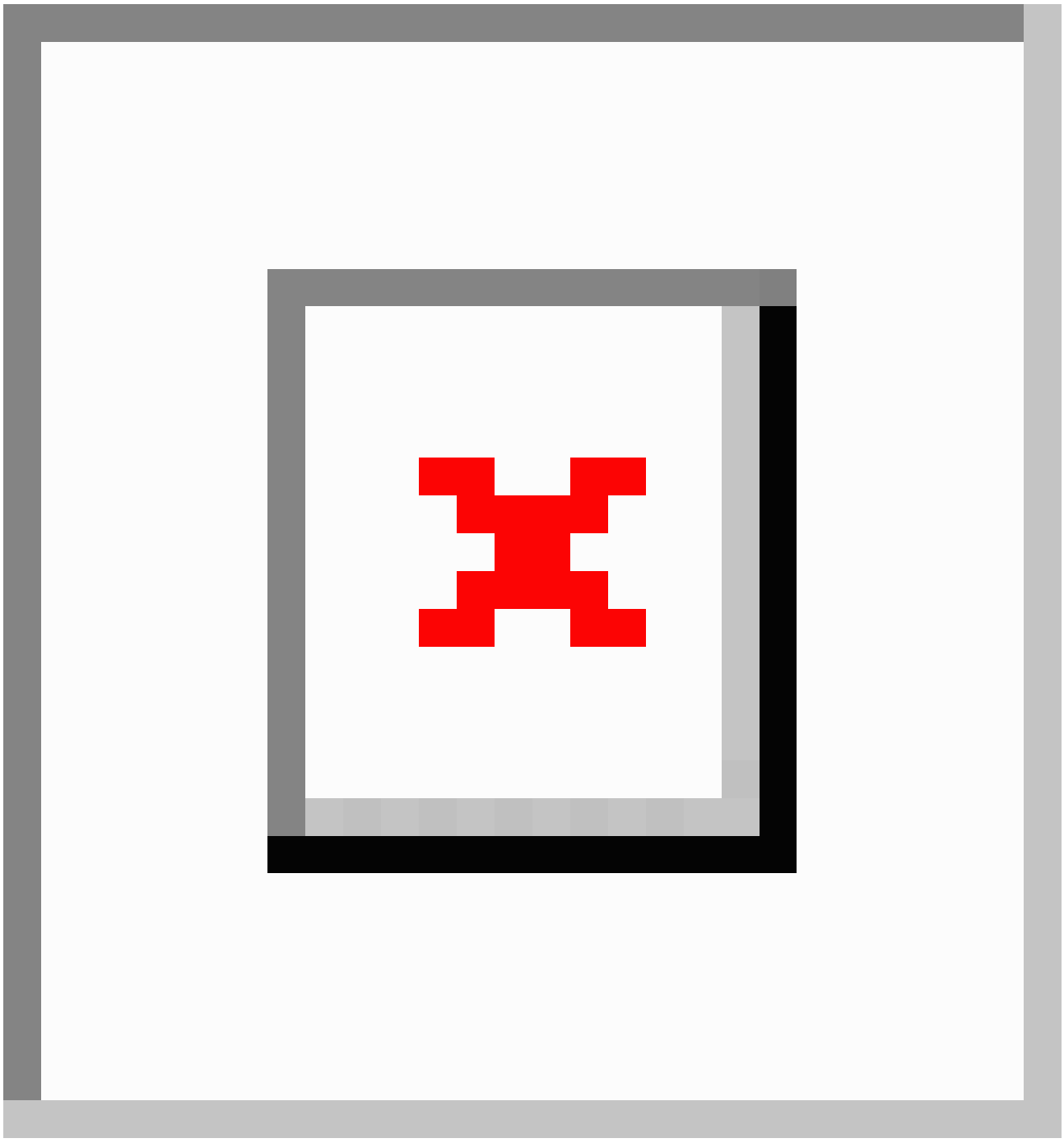


Mosaico for Designing, Performing, and Assessing Collaborative Clinical Cases

For the collaborative design of a clinical case, Mosaico provides standardized templates with all the information required. Students would write down relevant information about the clinical case, including age, sex, weight, height, physical exam, vital signs, laboratory tests, images, or videos (see Figure 3).

The web platform allows clinical cases to be generated, guided, and monitored; however, it stores them in a relational database. This does not allow records to be extracted and shared across heterogeneous technologies. In addition, Mosaico's lack of interoperability features limits the platform's use by forcing all institutions to use the same technology or make ad hoc integrations if they wish to benefit from the portfolio of clinical cases created.

Figure 3. Screenshot of the case design form provided by Mosaico, separated into the following sections: (A) patient information and vital signs, (**B**) patient's laboratory exams, and (C) images and videos (in this case, an x-ray of the thorax). HDL: high-density lipoprotein; LDL: low-density lipoprotein; VLDL: very low–density lipoprotein.



Extraction and Standardization of Clinical Cases

To provide Mosaico with interoperability capabilities, we proposed including a module that allows the transformation and export of clinical cases into a universal format—one that any other platform can easily read.

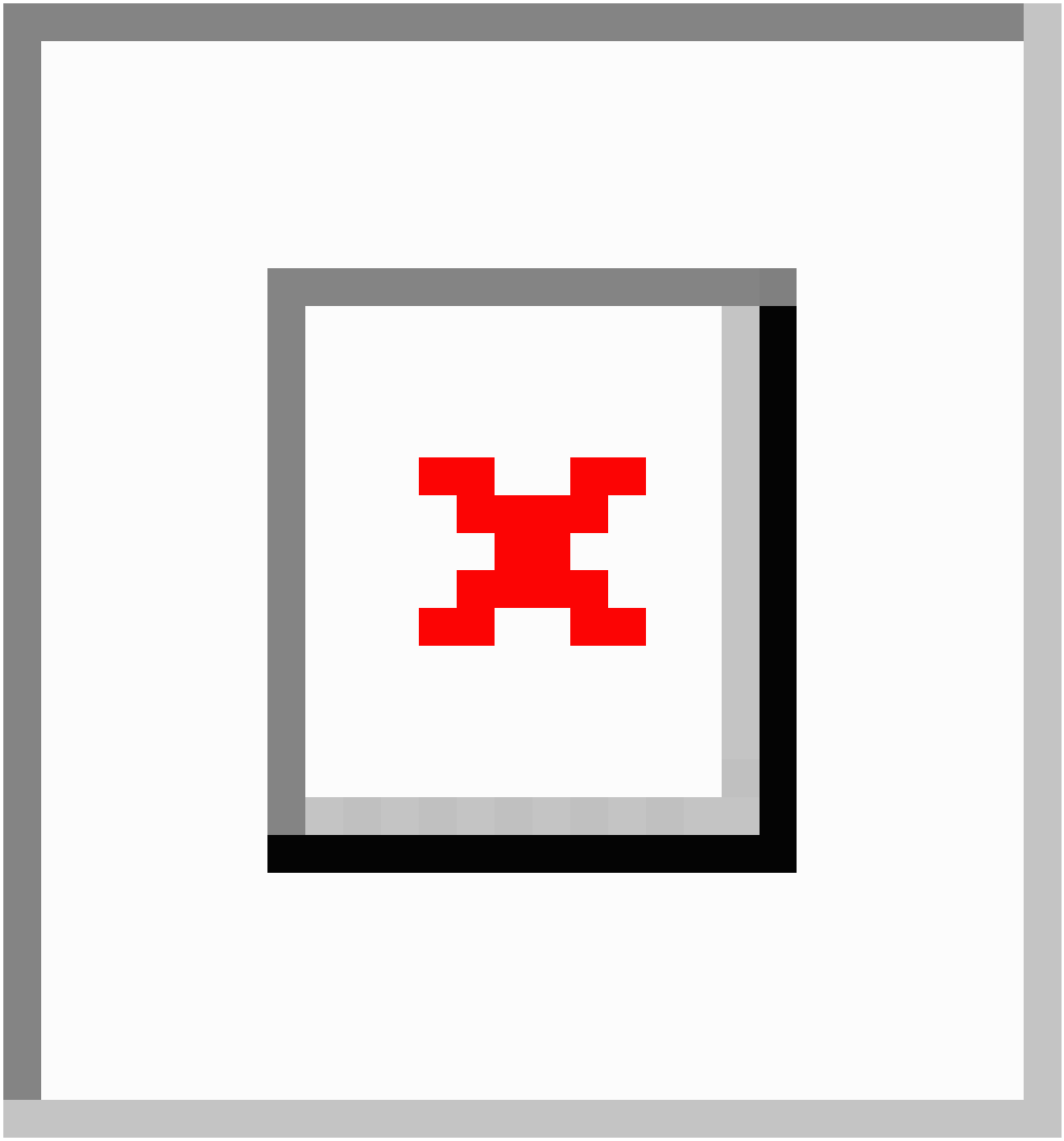
This module incorporates an interoperability layer designed and implemented through REST application programming interfaces [18] with HL7 FHIR [19] that offers clinical knowledge representation based on international standards. For this purpose, and according to the case design form provided by Mosaico (Figure 3), a set of HL7 FHIR resources would contain all the standardized clinical case information, making it easily exportable.

Transformation of Interoperable Clinical Cases With Clavy

Clavy is a learning object repository management platform based on formal grammars. Each Clavy repository comprises (1) a set of object schemata, which describes the structures of the different objects using formal grammars, and (2) the learning objects themselves. In addition, the platform fully supports Extract-Transform-Load (ETL) workflows with user-defined plug-ins and advanced editing facilities.

The tools integrated by Clavy can be accessed on the web and provide user-friendly interfaces to manage the repositories. Using these tools, domain experts without programming knowledge can use different transformations to reorganize the repository's structure, remove irrelevant information, and add new objects to this repository. For example, Figure 4 shows a screenshot that displays part of a specialized repository in the clinical domain, making it appear as a part of a schema-guided navigation tree, a part of the learning object collection, and a fragment of a particular learning object.

Figure 4. Clavy snapshot of a medical repository (navigation tree, learning object collection, and a particular learning object).



Results

Clinical Case

We applied our workflow ecosystem to generate standardized e-learning content for a relevant clinical case. The clinical case to show the proposed approach was related to a patient who consults for chest pain. The case was designed while considering all the clinical backgrounds necessary for its representation through a simulated patient. Furthermore, it was important to consider all the antecedents since the case will later be packaged and distributed to an LMS environment so that other students can consult it as study material.

Collaborative Chest Pain Case Development

Fourth-year medical students, along with an academic instructor from the University of Talca in Chile, developed the case in Mosaico by considering the following scenario:

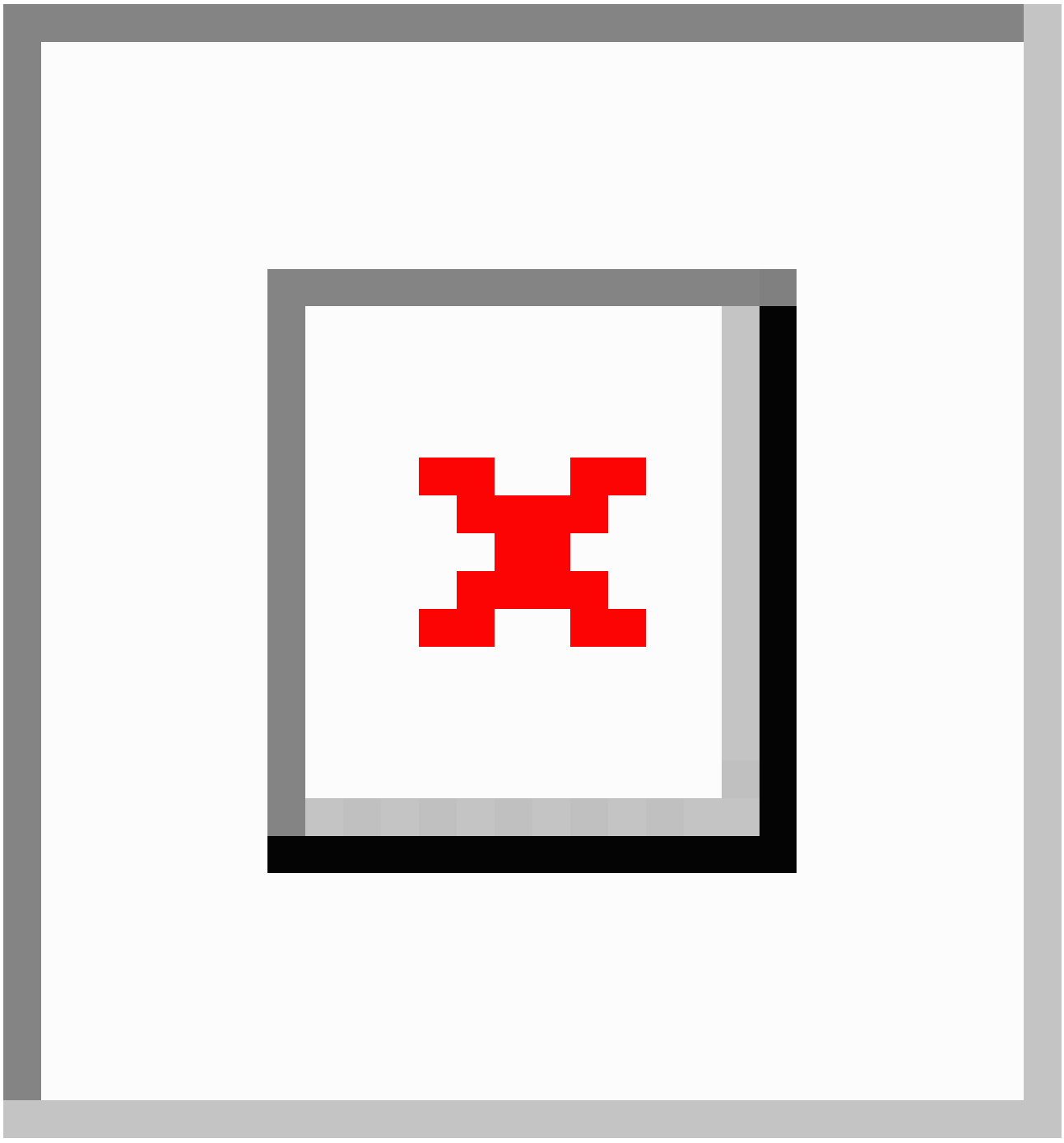
A 58-year-old male patient with a history of arterial hypertension (treated with 20 mg of enalapril every 12 h, which he takes irregularly), who works as an engineer 8 hours per day in an office, consults for oppressive chest pain in the precordial region, which is radiating to the neck and jaw, appearing at rest, and accompanied by coldness and cutaneous pallor that accentuated approximately 8 hours ago. He went to the hospital, was evaluated, and presented the following vital signs:

- Blood pressure: 170/115 mm Hg
- Heart rate: 95 beats per minute
- Respiratory rate: 22 respirations per minute
- Oxygen saturation: 98%
- Temperature: 36.7 °C

Moreover, the following complementary tests were incorporated: laboratory exams, chest x-rays, and an electrocardiogram taken in the hospital.

The case was created in Mosaico (see Figure 3) and stored in a MYSQL relational database; all data were then exported as HL7 FHIR resources. Figure 5 shows the resources involved, using a DiagnosticReport resource as the backbone of the case.

Figure 5. Chest pain case standardized with Health Level 7 (HL7) Fast Healthcare Interoperability Resources (FHIR) resources and respective JSON structure. EKG: electrocardiogram; Rx: prescription.

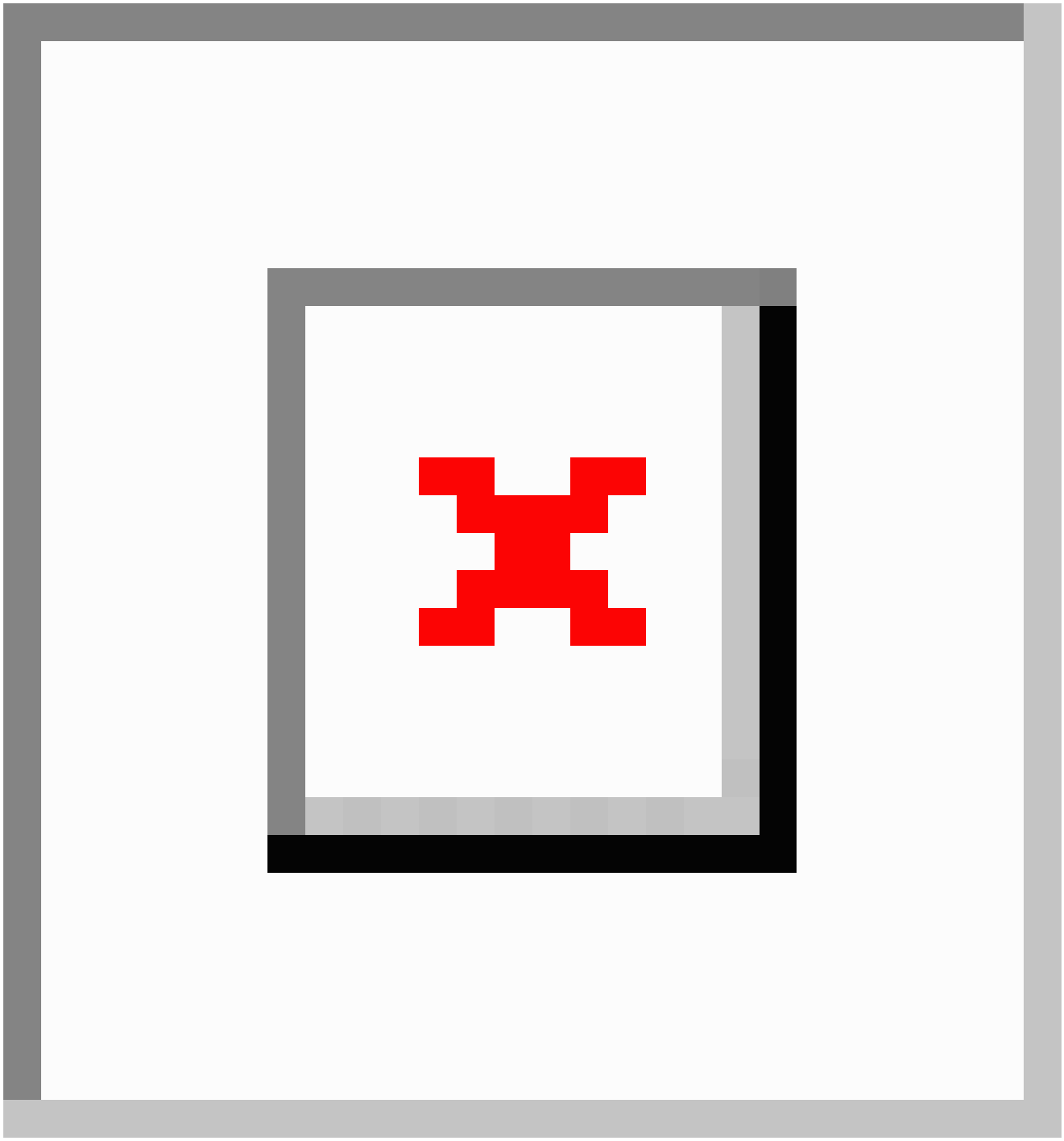


Use of Clavy to Generate Standardized e-Learning Content

Clavy was used to create an ETL pipeline to extract and transform the information created in Mosaico; this information could then be readily exported into IMS-CPs that could be deployed in an LMS, such as Moodle or Blackboard. In this way:

- The proposed pipeline begins with a generic JSON import plug-in, which lets Clavy ingest the Mosaico-generated JSON files encoding the HL7 FHIR entities associated with the clinical cases. As a result, a first repository is generated with (1) an object schema for each of the FHIR entity types ingested and (2) an object for each FHIR resource collected. Figure 6A shows a part of the Clavy schemata for the repository generated.
- Next, several automatic transformations are applied. The first transformation acts on the repository extracted by the JSON import plug-in to incorporate the FHIR semantics by making the relationships in FHIR instances explicit. The second transformation, in turn, enriches the repository with Systematized Nomenclature of Medicine (SNOMED) data. For this purpose, it detects SNOMED terms in learning objects, adds the SNOMED terms discovered to the repository as new objects, and links the FHIR entities with these new SNOMED objects as needed. Finally, the third transformation makes it possible to aggregate a set of interrelated FHIR entities into a single learning object. For this purpose, an entity that will act as the main one is selected, and a closure process is carried out to convert the nonhierarchical relationships into hierarchical ones. In this experience, the central entity chosen was the *Patient*.
- Finally, once a learning object repository focused on a particular kind of FHIR entity (*Patient*, in this case) is generated, its schema can be edited by using the Clavy repository editing facilities to accommodate all this information to the instructors' needs (in our case, the generation of standardized e-learning content). This experience allowed us to remove useless and irrelevant categories from the repository schema and, therefore, from the learning objects. Figure 6B illustrates the schema resulting from this editing step. Notice that, on the one hand, as produced by the previous transformation step, this schema characterizes only one type of learning object (*Patient*), aggregating all the relevant information in all the interrelated FHIR entities. However, on the other hand, many irrelevant FHIR information items from an educational point of view have been safely removed in this editing step.

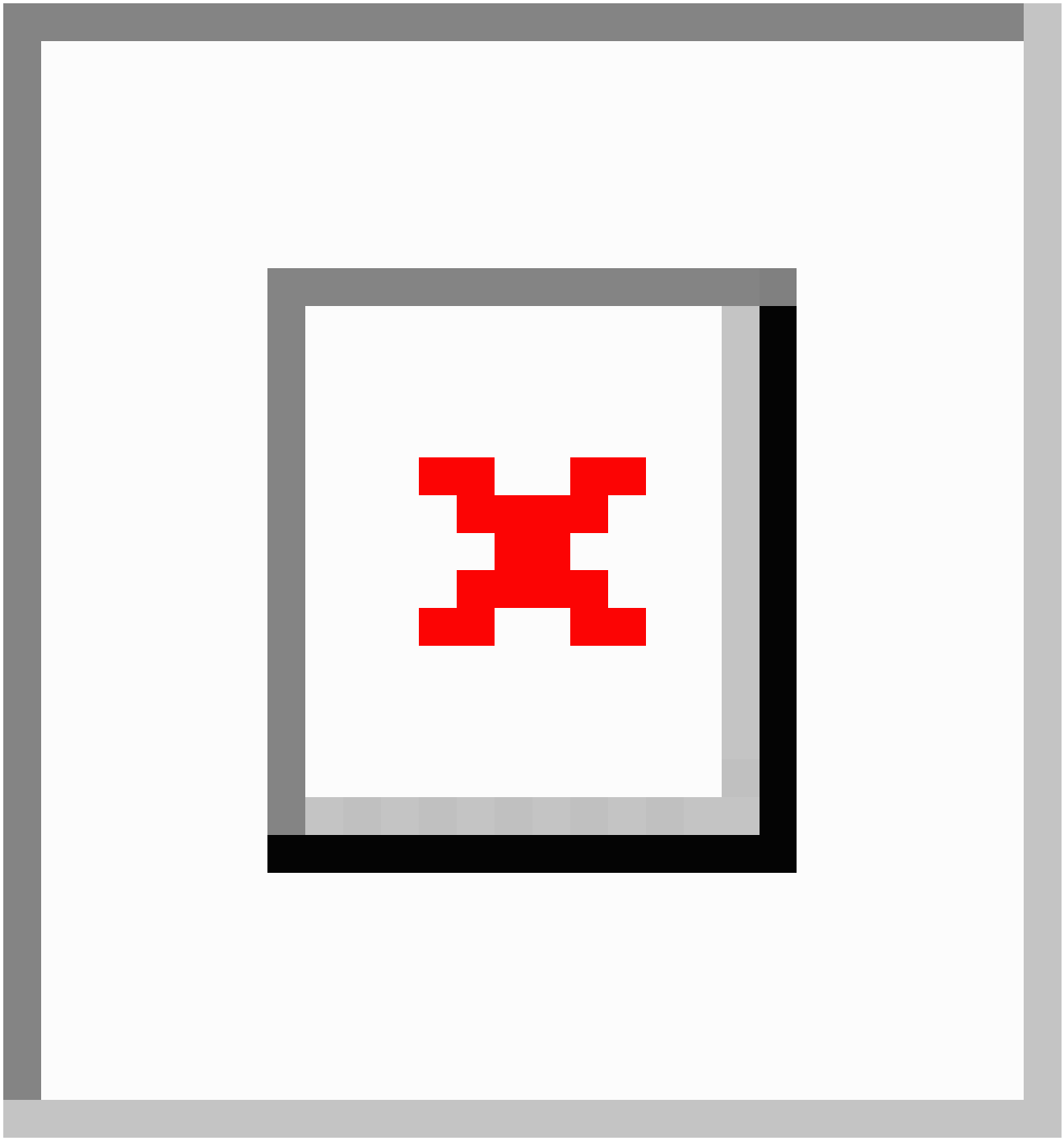
Figure 6. (A) Clavy schema for the Health Level 7 (HL7) Fast Healthcare Interoperability Resources (FHIR) entities generated by Mosaico and (**B**) the schema for the repository that results after the transformation and editing steps.



Deployment of IMS-CP in LMS Environments

Once the information has been transformed into a structure that suits the instructors' needs, a suitable Clavy export plug-in can be applied to the resulting repository to package the clinical cases into standardized learning packages that can be loaded into conventional LMSs. In this experience, the clinical case was exported as IMS-CPs and incorporated into Moodle (see Figure 7A). Figure 7B shows an IMS-CP manifest file segment associated with the original Mosaico clinical case.

Figure 7. (**A)** A learning object associated with the Mosaico clinical case imported into Moodle's platform as an Instructional Management Systems Content Package (IMS-CP) and (**B**) an excerpt of the IMS-CP manifest file.



Discussion

Principal Findings

In this paper, we identified four key elements to enable the effective use of clinical cases in e-learning environments: (1) the effective authoring of cases, (2) the interoperability of authoring tools, (3) the transformation of cases into standard e-learning content packages, and (4) the publication of content packages on widely used e-learning platforms. For this purpose, we proposed a workflow that consistently addresses each of these critical aspects by forming a technology ecosystem supported by (1) the Mosaico collaborative clinical case authoring tool, which supports students and teachers in health care careers with the collaborative design of clinical cases; (2) the HL7 FHIR clinical interoperability standard, which makes it possible to expose relevant information from the clinical cases to third-party tools; (3) the Clavy learning object repository management tool, which allows the transformation of these clinical cases into standardized educational formats; and (4) the Moodle LMS, which makes it possible to deploy the resulting contents for their consumption.

The collaborative design supported by Mosaico [7] allows medical students to create clinical cases through standardized templates with all the information required and the possibility to integrate relevant knowledge from multiple sources. The case study regarding chest pain described in the previous section, in

which Mosaico was successfully used to produce a complex clinical case by a group of advanced students under the supervision of a medical instructor, highlights the suitability of Mosaico in orchestrating this type of collaborative clinical case authoring. At this stage, it is essential to highlight Mosaico's ability to allow the creation of clinical cases by groups of students, which can be shared and exchanged democratically within the academic community. An interesting question to discuss is the wide variety of medical scenarios that can be considered during the design of clinical cases. Since the clinical cases are created by students in the company of an academic instructor, they are never repeated. Each design session has a new case history being shared, regardless of the diagnosis.

Concerning interoperability issues, HL7 FHIR specifications or notations [20] standardize medical records by incorporating an interoperable format that increases the exchange potential of any type of health data source. As exemplified by the case study in the previous section, the interoperability layer added to Mosaico highlights the feasibility of representing complex educational and clinical cases in terms of the HL7 FHIR information models. In this regard, clinical cases created using tools such as Mosaico, represented internally in proprietary and tool-dependent formats, can be offered to third-party tools in standardized and open formats. As shown in the case study, this aspect is essential to facilitating the reuse of clinical cases in many contexts that are not necessarily anticipated. In the work described in this paper, FHIR has allowed the generation of information packages used to build digital medical repositories, which can be deployed in subsequent stages of the design process. A similar environment to the one proposed and aimed at collecting clinical case information based on FHIR notations can be found in the Public Health Automated Case Event Reporting platform [21], which uses laboratory results to extend the description of clinical cases.

Although the availability of clinical cases represented in an interoperable format, such as HL7 FHIR, facilitates their consumption by external tools, a substantial gap prevents their publication in e-learning environments. Indeed, as evidenced in the case study presented, at the syntactic level, the HL7 FHIR information packages produced by Mosaico consist of JSON files and additional associated resources (eg, clinical images). Meanwhile, they represent instances of the FHIR information models at the semantic level. In contrast, information publishable on e-learning platforms follows well-established e-learning standards (eg, IMS-CP or SCORM). Therefore, this is a domain transformation problem at the syntactic level (eg, the transformation of JSON documents into IMS-CP XML manifests) and at the semantic level (eg, the change of FHIR entity graphs into IMS-CP or SCORM organizations). As the case study shows, the Clavy platform demonstrated its ability to undertake the required transformation tasks. Indeed, in this paper, Clavy was able to convert the HL7 FHIR resources produced by Mosaico through automatic procedures, taking the resources as input and generating digital repositories to be edited and processed. In addition, during this transformation stage, the remarkable functionality of Clavy allowed text elements to be converted from Mosaico-based clinical cases into standard terms such as SNOMED Clinical Terms or Logical Observation Identifiers Names and Codes (LOINC), which adds value to the original case descriptions. The advantages of using semantic standardization terminologies have been reviewed in matters such as structured reporting [22] or health care process mining [23].

The final stage consists of generating standard educational specifications, which can be useful when spreading instructional content across diverse health care e-learning environments [24]. In this case, Clavy allowed the export of the clinical cases collected from Mosaico using standard formats such as IMS-CP or SCORM. Standard specifications have played a crucial role in applying e-learning technologies for medical education. Most of these specifications focus on the content sharing of health care learning resources that can be designed and developed collaboratively, as seen in the mEducator initiative [25]. The work presented in this paper is similar to this initiative, relying on learning objects and IMS-CP educational packages to deliver instructional clinical cases to medical students and health care practitioners. In addition, the role of technical standards in medical instructional settings has been reviewed as a part of recent digital innovations introduced in health care education and training [26]. These innovations boost the use of SCORM packages that allow e-learning developers to add enhanced interactive features and assessment procedures to the massive amount of medical knowledge available, which are traditionally stored as plain content items [27]. Moreover, they introduce the possibility of tracking user interactions with these content items to supervise their learning in clinical scenarios [28].

Conclusions

In this paper, we proposed and implemented a workflow to deploy standard e-learning content, combining Mosaico, a tool for collaborative content authoring; HL7 FHIR, a clinical information interoperability standard; and Clavy, a platform for the flexible management of learning object repositories. Through a substantial case study, we demonstrated (1) how clinical cases collaboratively developed in Mosaico can be coded and exported in HL7 FHIR, (2) how this facilitates the reuse of these clinical cases with other tools, and (3) how cases exported in HL7 FHIR can be transformed into IMS-CP e-learning packages using a tool such as Clavy. As a result, we demonstrated the feasibility of deploying collaboratively created content using proprietary authoring tools in conventional open e-learning environments.

Having demonstrated the approach's feasibility, we intend to apply it to other scenarios of collaborative production of medical e-learning content. Likewise, we want to better integrate the components that constitute the proposed workflow ecosystem to the end users transparently, thus increasing usability and the overall user experience.

Acknowledgments

This research is supported by Research Project PID2021-123048NB-I00 and Research Project CTI220001 (CENS).

Conflicts of Interest

None declared.

References

1. Cucinotta D, Vanelli M. WHO declares COVID-19 a pandemic. Acta Biomed 2020 Mar 19;91(1):157-160. [doi: 10.23750/abm.v91i1.9397] [Medline: 32191675]
2. Monaghan AM. Medical teaching and assessment in the era of COVID-19. J Med Educ Curric Dev 2020 Oct 16;7:2382120520965255. [doi: 10.1177/2382120520965255] [Medline: 33117891]
3. Rajab MH, Gazal AM, Alkattan K. Challenges to online medical education during the COVID-19 pandemic. Cureus 2020 Jul 2;12(7):e8966. [doi: 10.7759/cureus.8966] [Medline: 32766008]
4. Inzunza M, Besser N, Bellolio F. Decrease in operative volume in general surgery residents in Chile: effects of the COVID-19 pandemic. Br J Surg 2021 Jun 22;108(6):e226-e227. [doi: 10.1093/bjs/znab082] [Medline: 33760034]
5. Papapanou M, Routsi E, Tsamakis K, Fotis L, Marinos G, Lidoriki I, et al. Medical education challenges and innovations during COVID-19 pandemic. Postgrad Med J 2022 May;98(1159):321-327. [doi: 10.1136/postgradmedj-2021-140032] [Medline: 33782202]
6. Zuo L, Dillman D, Miller Juvé A. Learning at home during COVID-19: a multi-institutional virtual learning collaboration. Med Educ 2020 Jul;54(7):664-665. [doi: 10.1111/medu.14194] [Medline: 32330317]
7. Guinez-Molinos S, Gonzalez Díaz J, Gomar Sancho C, Espinoza P, Constenla G. A web platform (MOSAICO) to design, perform, and assess collaborative clinical scenarios for medical students: viewpoint. JMIR Med Educ 2021 Jan 26;7(1):e23370. [doi: 10.2196/23370] [Medline: 33496676]
8. Haynes MR, Gaglani SM, Wilcox MV, Mitchell T, DeLeon V, Goldberg H. Learning through Osmosis: a collaborative platform for medical education. Innovations in Global Medical and Health Education 2014 Nov;2014(1). [doi: 10.5339/igmhe.2014.2]
9. Cabrera D, Cooney R. Wikis: using collaborative platforms in graduate medical education. J Grad Med Educ 2016 Feb;8(1):99-100. [doi: 10.4300/JGME-D-15-00567.1] [Medline: 26913112]
10. Vicente CR, Jacobs F, de Carvalho DS, Chhaganlal K, de Carvalho RB, Raboni SM, et al. The Joint Initiative for Teaching and Learning on Global Health Challenges and One Health experience on implementing an online collaborative course. One Health 2022 Jul 17;15:100409. [doi: 10.1016/j.onehlt.2022.100409] [Medline: 36277091]
11. Quintana Y, Einstein D, Joyce R, Lyu A, El Sayed N, Andrews C, et al. Accelerating learning health systems using Alicanto collaboration platforms. JCO Glob Oncol 2022 May;8(Supplement_1):62. [doi: 10.1200/GO.22.68000]
12. Turnbull D, Chugh R, Luck J. Learning management systems, an overview. In: Tatnall A, editor. Encyclopedia of Education and Information Technologies. Cham, Switzerland: Springer; 2020:1052-1058. [doi: 10.1007/978-3-030-10576-1]
13. Guinez-Molinos S, Martínez-Molina A, Gomar-Sancho C, Arias González VB, Szyld D, García Garrido E, et al. A collaborative clinical simulation model for the development of competencies by medical students. Med Teach 2017 Feb;39(2):195-202. [doi: 10.1080/0142159X.2016.1248913] [Medline: 27841066]
14. Guinez-Molinos S, Maragaño Lizama P, Gomar-Sancho C. Collaborative clinical simulation to train medical students. Article in Spanish. Rev Med Chil 2018 May;146(5):643-652. [doi: 10.4067/s0034-98872018000500643] [Medline: 30148928]
15. Benson T, Grieve G. Principles of Health Interoperability: SNOMED CT, HL7 and FHIR. Cham, Switzerland: Springer; 2016. [doi: 10.1007/978-3-319-30370-3]
16. Gayoso-Cabada J, Gomez-Albarran M, Sierra JL. Enhancing the browsing cache management in the Clavy platform. Presented at: 2018 International Symposium on Computers in Education (SIIE); September 19-21, 2018; Jerez, Spain p. 1-6. [doi: 10.1109/SIIE.2018.8586778]
17. Athaya H, Nadir RDA, Indra Sensuse D, Kautsarina K, Suryono RR. Moodle implementation for E-learning: a systematic review. Presented at: SIET '21: 6th International Conference on Sustainable Information Engineering and Technology 2021; September 13-14, 2021; Malang, Indonesia p. 106-112. [doi: 10.1145/3479645.3479646]
18. Fielding RT. Architectural styles and the design of network-based software architectures [Dissertation]. University of California, Irvine. 2000. URL: www.ics.uci.edu/~fielding/pubs/dissertation/top.htm [accessed 2023-08-17]
19. Ayaz M, Pasha MF, Alzahrani MY, Budiarto R, Stiawan D. The Fast Health Interoperability Resources (FHIR) standard: systematic literature review of implementations, applications, challenges and opportunities. JMIR Med Inform 2021 Jul 30;9(7):e21929. [doi: 10.2196/21929] [Medline: 34403353]
20. Bender D, Sartipi K. HL7 FHIR: an agile and RESTful approach to healthcare information exchange. Presented at: CBMS 2013 - 26th IEEE International Symposium on Computer-Based Medical Systems; June 20-22, 2013; Porto, Portugal p. 326-331. [doi: 10.1109/CBMS.2013.6627810]

21. Mishra N, Duke J, Karki S, Choi M, Riley M, Ilatovskiy AV, et al. A modified public health automated case event reporting platform for enhancing electronic laboratory reports with clinical data: design and implementation study. J Med Internet Res 2021 Aug 11;23(8):e26388. [doi: 10.2196/26388] [Medline: 34383669]
22. Fennelly O, Grogan L, Reed A, Hardiker NR. Use of standardized terminologies in clinical practice: a scoping review. Int J Med Inform 2021 May;149:104431. [doi: 10.1016/j.ijmedinf.2021.104431] [Medline: 33713915]
23. Helm E, Lin AM, Baumgartner D, Lin AC, Küng J. Towards the use of standardized terms in clinical case studies for process mining in healthcare. Int J Environ Res Public Health 2020 Feb 19;17(4):1348. [doi: 10.3390/ijerph17041348] [Medline: 32093073]
24. Buendía F, Gayoso-Cabada J, Sierra JL. Generation of standardized e-learning content from digital medical collections. J Med Syst 2019 May 18;43(7):188. [doi: 10.1007/s10916-019-1330-5] [Medline: 31104150]
25. Konstantinidis S, Kaldoudi E, Bamidis PD. Enabling content sharing in contemporary medical education: a review of technical standards. The Journal on Information Technology in Healthcare 2009 Dec;7(6):363-375.
26. Komenda M, Karolyi M, Woodham L, Vaitsis C. Chapter four: the role of technical standards in healthcare education. In: Digital Innovations in Healthcare Education and Training. London, United Kingdom: Academic Press; Jan 2021:47-59. [doi: 10.1016/B978-0-12-813144-2.00004-0]
27. Dutkiewicz A, Kołodziejczak B, Leszczyński P, Mokwa-Tarnowska I, Topol P, Kupczyk B, et al. Online interactivity – a shift towards e-textbook-based medical education. Studies in Logic, Grammar and Rhetoric 2018 Dec 1;56(1):177-192. [doi: 10.2478/slgr-2018-0048]
28. Mangina E, McGirl T, Ryan G, Murphy J, McAuliffe F. Experience API (xAPI) for virtual reality (VR) education in medicine. In: Auer ME, Pester A, May D, editors. Learning With Technologies and Technologies in Learning. Cham, Switzerland: Springer; 2022:335-359. [doi: 10.1007/978-3-031-04286-7_16]

Abbreviations

ETL: Extract-Transform-Load
FHIR: Fast Healthcare Interoperability Resources
HL7: Health Level 7
IMS-CP: Instructional Management Systems Content Packaging
LMS: learning management system
LOINC: Logical Observation Identifiers Names and Codes
SCORM: Sharable Content Object Reference Model
SNOMED: Systematized Nomenclature of Medicine
WHO: World Health Organization

Edited by C Lovis; submitted 23.12.22; peer-reviewed by K Gupta, L Lopez, N Jiwani; revised version received 31.07.23; accepted 02.08.23; published 19.09.23.

Please cite as:
Guinez-Molinos S, Buendía-García F, Sierra-Rodríguez JL, Gayoso-Cabada J, González-Díaz J
Computer-Supported Collaborative Design of Standardized Clinical Cases: Algorithm Development and Validation
JMIR Med Inform 2023;11:e45315
URL: https://medinform.jmir.org/2023/1/e45315
doi:10.2196/45315

Assessing the Use of German Claims Data Vocabularies for Research in the Observational Medical Outcomes Partnership Common Data Model: Development and Evaluation Study

Elisa Henke[1], MSc; Michéle Zoch[1], Dipl Wi Inf; Michael Kallfelz[2], Dr med; Thomas Ruhnke[3], Dipl Wi Math; Liz Annika Leutner[1], MSc; Melissa Spoden[3], DrPH; Christian Günster[3], Dipl Math; Martin Sedlmayr[1], Dr rer nat, Prof Dr; Franziska Bathelt[4], Dr rer nat

1

2

3

4

Corresponding Author:
Elisa Henke, MSc

Abstract

Background: National classifications and terminologies already routinely used for documentation within patient care settings enable the unambiguous representation of clinical information. However, the diversity of different vocabularies across health care institutions and countries is a barrier to achieving semantic interoperability and exchanging data across sites. The Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) enables the standardization of structure and medical terminology. It allows the mapping of national vocabularies into so-called standard concepts, representing normative expressions for international analyses and research. Within our project "Hybrid Quality Indicators Using Machine Learning Methods" (Hybrid-QI), we aim to harmonize source codes used in German claims data vocabularies that are currently unavailable in the OMOP CDM.

Objective: This study aims to increase the coverage of German vocabularies in the OMOP CDM. We aim to completely transform the source codes used in German claims data into the OMOP CDM without data loss and make German claims data usable for OMOP CDM–based research.

Methods: To prepare the missing German vocabularies for the OMOP CDM, we defined a vocabulary preparation approach consisting of the identification of all codes of the corresponding vocabularies, their assembly into machine-readable tables, and the translation of German designations into English. Furthermore, we used 2 proposed approaches for OMOP-compliant vocabulary preparation: the mapping to standard concepts using the Observational Health Data Sciences and Informatics (OHDSI) tool Usagi and the preparation of new 2-billion concepts (ie, *concept_id* >2 billion). Finally, we evaluated the prepared vocabularies regarding completeness and correctness using synthetic German claims data and calculated the coverage of German claims data vocabularies in the OMOP CDM.

Results: Our vocabulary preparation approach was able to map 3 missing German vocabularies to standard concepts and prepare 8 vocabularies as new 2-billion concepts. The completeness evaluation showed that the prepared vocabularies cover 44.3% (3288/7417) of the source codes contained in German claims data. The correctness evaluation revealed that the specified validity periods in the OMOP CDM are compliant for the majority (705,531/706,032, 99.9%) of source codes and associated dates in German claims data. The calculation of the vocabulary coverage showed a noticeable decrease of missing vocabularies from 55% (11/20) to 10% (2/20) due to our preparation approach.

Conclusions: By preparing 10 vocabularies, we showed that our approach is applicable to any type of vocabulary used in a source data set. The prepared vocabularies are currently limited to German vocabularies, which can only be used in national OMOP CDM research projects, because the mapping of new 2-billion concepts to standard concepts is missing. To participate in international OHDSI network studies with German claims data, future work is required to map the prepared 2-billion concepts to standard concepts.

(JMIR Med Inform 2023;11:e47959) doi:10.2196/47959

KEYWORDS

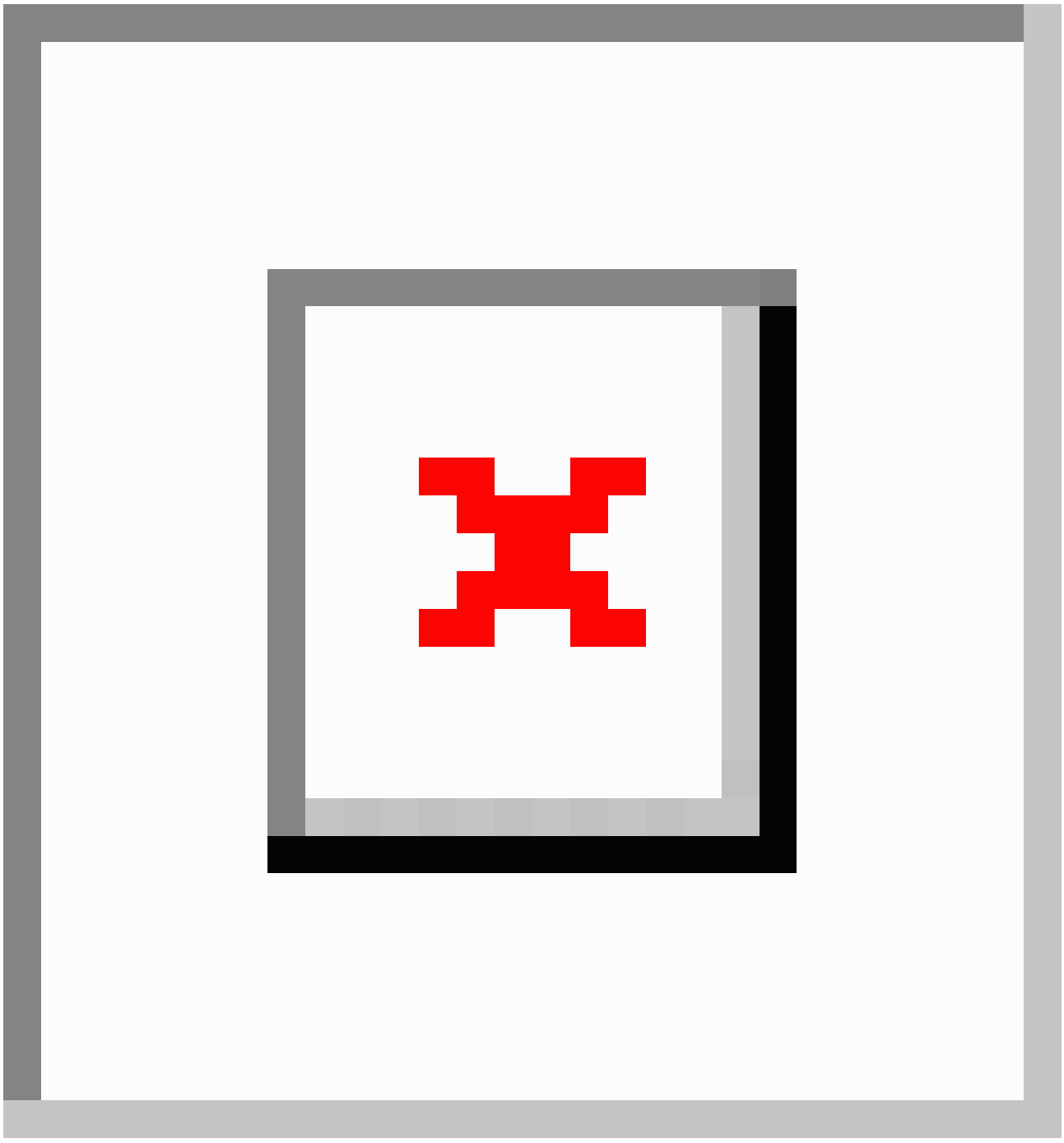
OMOP CDM; interoperability; vocabularies; claims data; OHDSI; Observational Medical Outcomes Partnership; common data model; Observational Health Data Sciences and Informatics

Introduction

Background and Significance

To generate reliable evidence in the health care sector, real-world data (RWD) can be used. RWD comprise observational data that are routinely collected in the context of patient care from various sources [1]. National classifications and terminologies enable the unambiguous representation of, for example, diagnoses (*International Classification of Diseases, Tenth Revision, German Edition* [*ICD-10-GM*]), procedures (*Operationen- und Prozedurenschlüssel* [Operations and Procedures Classification]), or laboratory data (Logical Observation Identifiers Names and Codes). Because these vocabularies are already routinely used for documentation within patient care settings, RWD already have a structured set of clinical information. However, each health care institution and country can have their own classifications, terminologies, or internally used set of codes. The diversity of different vocabularies is a barrier to achieving semantic interoperability and exchanging data across health care institutions and countries, as exemplarily shown in Figure 1. The code "C03" has 5 different semantic meanings covering the domains of drug, anatomic site, procedure, and condition [2]. Conducting research based on local vocabularies, terminologies, or classifications would result in custom analysis scripts for each site involved in a study. This not only entails high maintenance and time costs but is also unsustainable.

Figure 1. Overview of different meanings of the code C03 across various vocabularies. CAMS: Chinese Academy of Medical Sciences; CMS: Centers for Medicare & Medicaid Services; NCHS: National Center for Health Statistics; NHS: National Health Service; OPCS: Office of Population Censuses and Surveys; WHO: World Health Organization.



As a prerequisite for using data from heterogeneous data sources for international research and thus preventing the development of individual analysis scripts, harmonization and transformation to a common data model (CDM) are required. In recent years, the Observational Medical Outcomes Partnership (OMOP) CDM fostered by the Observational Health Data Sciences and Informatics (OHDSI) has become an essential open community data standard for research with RWD [3,4]. The OMOP CDM combines standardized data tables with centrally provided standardized vocabularies to ensure syntactic and semantic interoperability. The standardized vocabularies are represented in the OMOP CDM through concepts that enable the unique identification of all clinical events in the OMOP CDM. The concepts in the OMOP CDM are divided into standard and nonstandard concepts. Standard concepts provide normative expressions for international analyses and research based on the OMOP CDM. As an example, the Standard Nomenclature of Medicine concepts are mostly standard concepts in the OMOP CDM, for example, for the *Condition* domain. In contrast, nonstandard concepts are used to store codes of national vocabularies often used in source data, such as *ICD-10-GM*. The conversion ("mapping") of nonstandard to standard concepts is part of OMOP CDM vocabulary tables and represented as concept relationships. Similar to the concepts themselves, the mapping is provided through the central OHDSI vocabularies repository Athena [5]. The advantage of standardized OMOP

tables is that the source values and associated concepts, as well as the standardized concepts, are retained. Thus, it is always clear (1) what was part of the source data and (2) which concepts can be used for international research.

The main challenge faced by many researchers is the mapping of local source codes to OMOP CDM standard concepts [6-9]. Within our project "Hybrid Quality Indicators Using Machine Learning Methods" (Hybrid-QI) [10], we face this challenge during the transformation of German clinical data and claims data into the OMOP CDM. The aim of the project is the linkage of German clinical data and claims data into the OMOP CDM, to increase the effectiveness of quality measurement based on risk-adjusted quality indicators. To enable research based on 2 different heterogeneous data sets, the OMOP CDM is used for data harmonization. Although we have already successfully completed the semantic mapping of our clinical data [11], the mapping of German claims data comprising inpatient and outpatient data is still an open issue. To address this challenge, we focus first on the vocabularies used in German claims data as they build the basis for further semantic mapping to the OMOP CDM.

State of the Art

In an initial analysis of the current coverage of German claims data vocabularies in the OMOP CDM by Henke et al [12], it was shown that 55% (11/20) of the vocabularies are not available. Only 15% (3/20) of the vocabularies are currently present in Athena. The remaining 30% (6/20) of the vocabularies can be mapped to standard concepts in the OMOP CDM by using the *source_to_concept_map* table in the OMOP CDM. As a consequence, not all source codes used in German claims data can be represented in the OMOP CDM using standard concepts. Instead of losing the data during the transformation into the OMOP CDM, Sathappan et al [13] suggest storing the source codes in the **_source_value* columns of the OMOP CDM and mapping them to the *concept_id* of 0. However, this would again lead to the problem of using source codes for research as described earlier. In particular, the lack of associated nonstandard concepts in the OMOP CDM results in a loss of information about the vocabulary originally used in the source.

To our knowledge, no one so far has tried to prepare the missing 55% (11/20) of the German vocabularies for the OMOP CDM.

Objectives

The purpose of this paper is to increase the coverage of German vocabularies in the OMOP CDM. With our study, we want to make a major contribution to research with RWD based on the OMOP CDM by:

1. Completely transforming source codes used in German claims data to the OMOP CDM without data loss and
2. Making German claims data usable for research based on the OMOP CDM.

Methods

Ethical Considerations

Only synthesized claims data were used for our purposes, and therefore, no ethics approval was required.

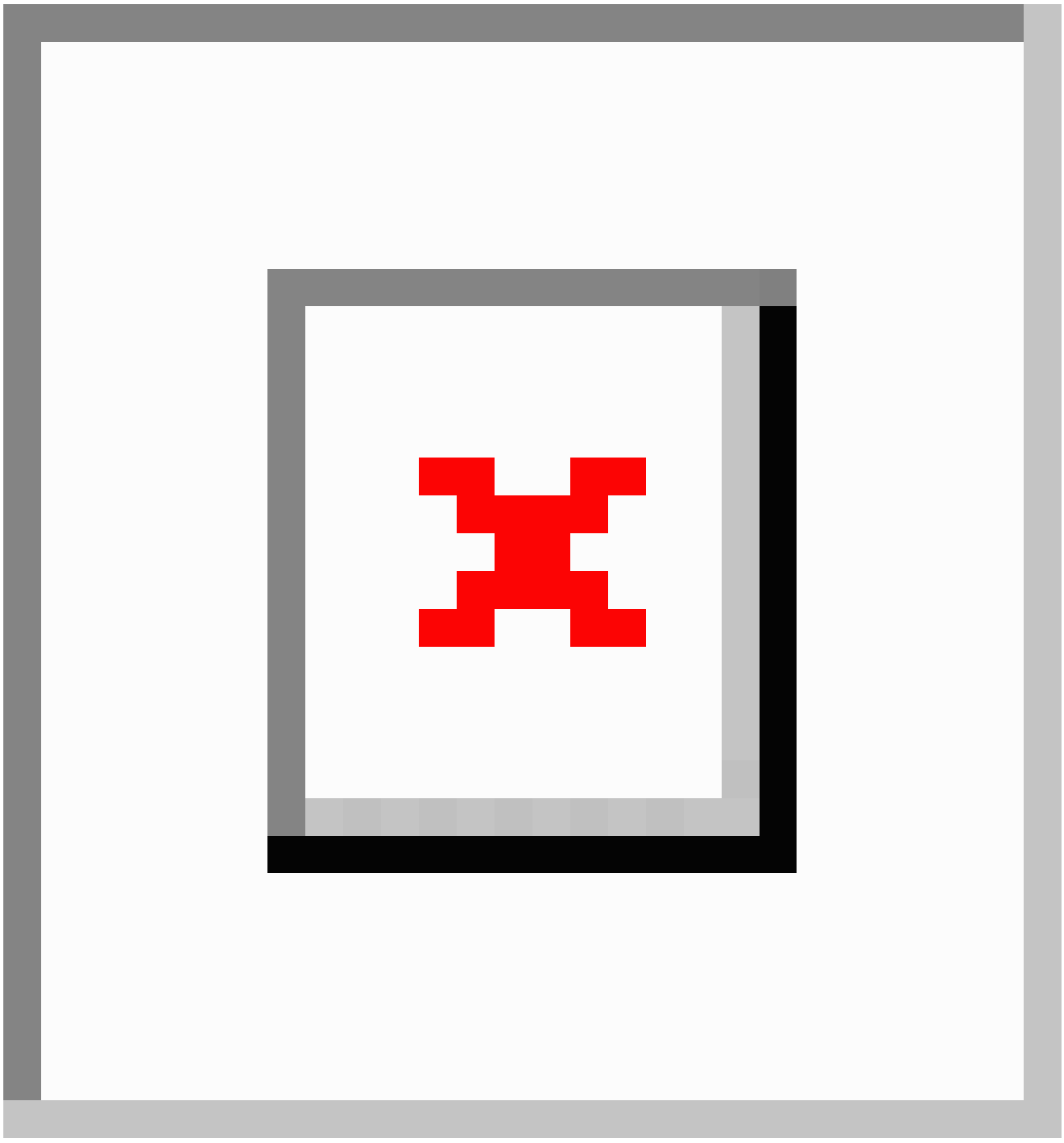
Study Data

The preparation of missing German vocabularies and its evaluation was done in the context of the Hybrid-QI project. The following four different medical indications were chosen as examples for the development of hybrid quality indicators:

1. Acute myocardial infarction
2. Cerebral infarction or intracerebral hemorrhage
3. Colorectal resection for carcioma
4. Shoulder endoprosthesis or osteosynthesis for proximal humerus fracture

To provide source data for data harmonization in the OMOP CDM, we used synthesized claims data from the German local health care funds (Allgemeine Ortskrankenkassen), which are based on real data. This data set includes billing data from 10,000 patients for a 6-year period comprising 558 MB of tabular data. For our purpose, we only focused on the source codes of the missing German vocabularies and their associated documentation or billing date. The German vocabularies considered in the following steps for the preparation in the OMOP CDM are summarized in Figure 2. We categorized the vocabularies according to their definitions and codes as remedy, billing, drug, condition, and provider vocabularies.

Figure 2. German vocabularies currently not available in the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM). ASV: Ambulante spezialfachärztliche Versorgung; EBM: Einheitlicher Bewertungsmaßstab; HMK: Heilmittelkatalog; HPNR: Heilmittelpositionsnummern; PIA: Bundeseinheitlicher Katalog für die Dokumentation der Leistungen der psychiatrischen Institutsambulanzen; PZN: Pharmazentralnummer.



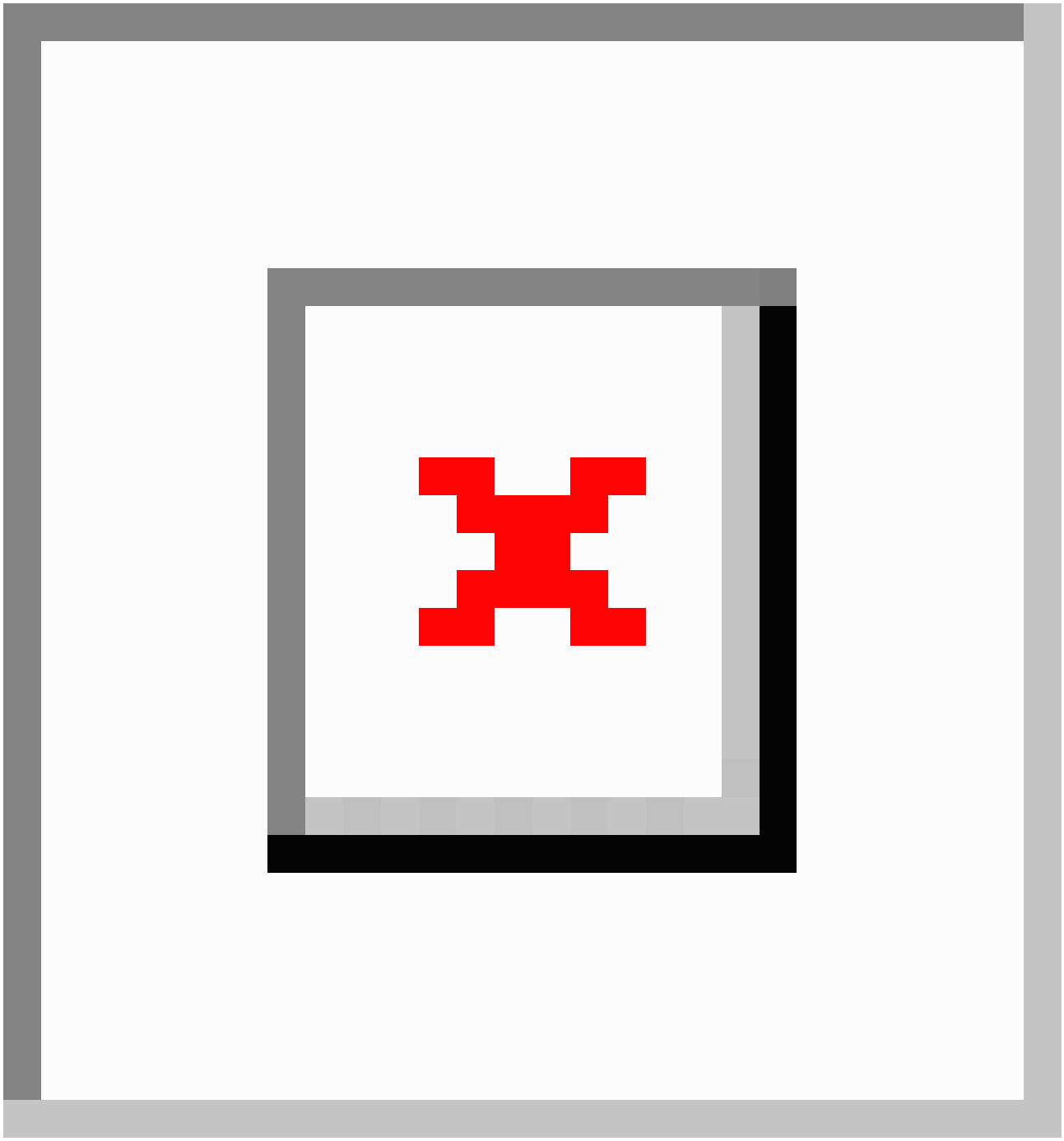
Vocabulary Preparation Approach

Overview

To prepare missing German vocabularies for the OMOP CDM, we defined an approach divided into 5 main steps (see Figure 3). First, we performed a selective search to identify all codes of the corresponding vocabulary. We especially checked the license restrictions of the vocabularies that do not permit their preparation for the OMOP CDM. Based on the search results from the first step, we summarized all codes, the hierarchy of the vocabulary, the corresponding German terms, and validity periods per vocabulary in tabular form for those vocabularies that do not have license restrictions. If no validity period was found, we used the default OMOP CDM values for *valid_start_date* (“1970-01-01”) and *valid_end_date* (“2099-12-31”) [14]. Next, we translated the German designations into English to make the semantic meaning of the codes understandable in an international context. Two researchers with domain knowledge were involved in the translations and their validation. For the fourth step of OMOP-compliant vocabulary preparation, there were 2 different methods available [15]: mapping to standard concepts using Usagi [16,17] and OMOP-compliant preparation of new 2-billion concepts (ie, *concept_id* >2 billion). Both methods are

described in more detail in the following sections. Finally, we evaluated the results from the fourth step.

Figure 3. Vocabulary preparation approach. OMOP: Observational Medical Outcomes Partnership.



Mapping to Standard Concepts

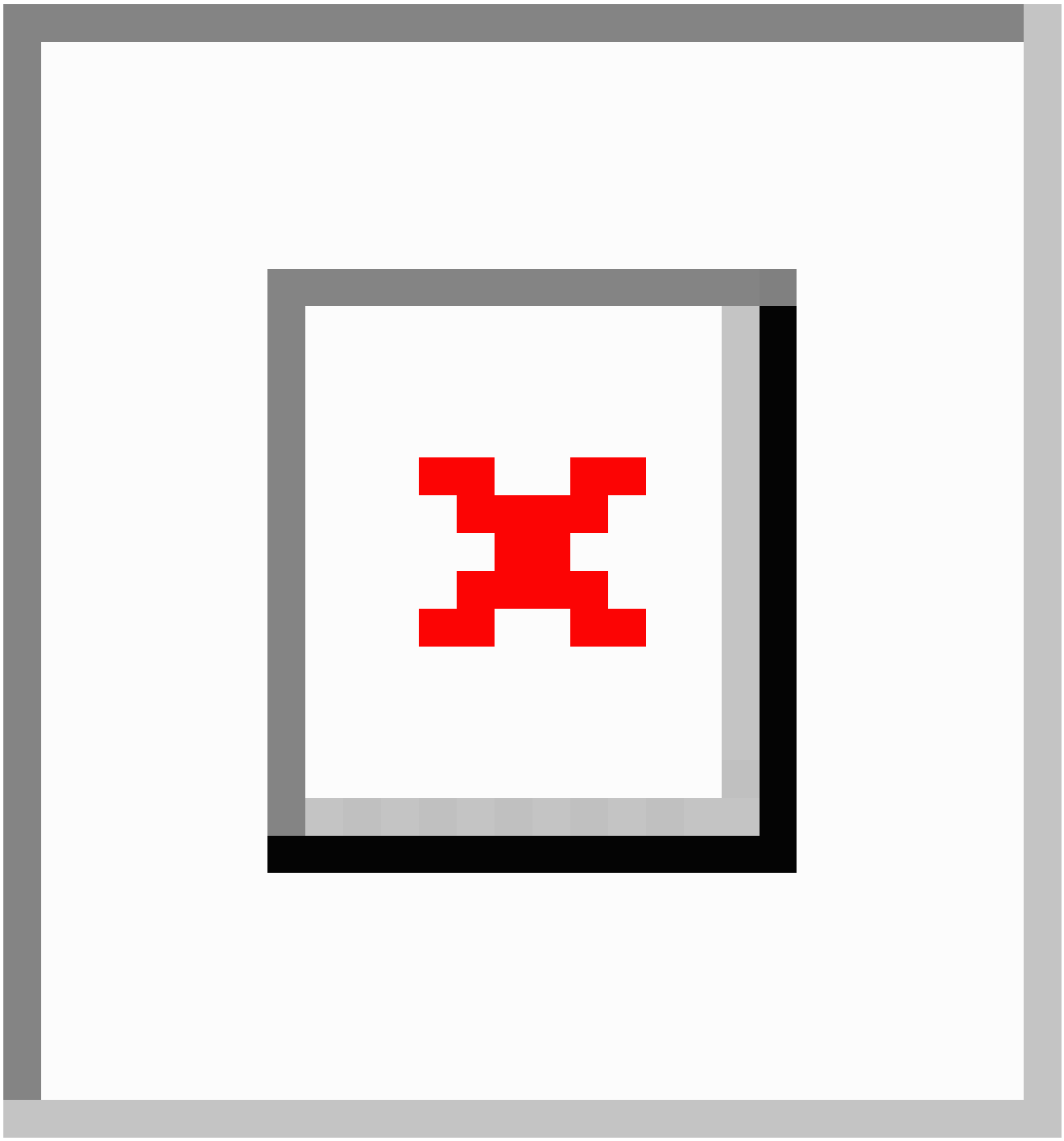
To map local codes from a source system to standard concepts used in the OMOP CDM, the OHDSI community provides the open-source tool Usagi [16,17]. Usagi uses a term-similarity approach to propose appropriate standard concepts in the OMOP standardized vocabulary based on the English designation of the source codes. In the first step, we loaded a prepared list of source codes and their German and English designations into Usagi. Next, we specified the target domain of the OMOP standardized vocabulary that should be used during the mapping, based on the categories assigned to the German vocabularies shown in Figure 2. In a team of 2 researchers with domain knowledge, we reviewed the proposals made by Usagi and jointly discussed and resolved conflicts that had arisen during the mapping process. After the 2 researchers approved all proposals, the mapping of the source codes to standard concepts was exported as a CSV file following the format structure of the OMOP CDM *source_to_concept_map* table.

New 2-Billion Concepts

If source codes cannot be mapped using the standardized OMOP vocabularies, it is possible to create new 2-billion concepts for the OMOP CDM [15], which is a number range reserved for local *concept* creation. With this approach, we prepared the missing vocabularies to be OMOP compliant for the OMOP

CDM tables *vocabulary*, *concept*, *concept_class*, and *concept_relationship*. Figure 4 shows the preparation of a new 2-billion concept for the OMOP CDM *concept* table for the *Heilmittelpositionsnummern* (HPNR; uniform list of item numbers for therapeutic services) source code "1514."

Figure 4. A new 2-billion concept in the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) *concept* table for HPNR source code "1514." HPNR: Heilmittelpositionsnummern (uniform list of item numbers for therapeutic services).



During the preparation process, it must be taken into account that the new concepts are assigned a *concept_id* >2 billion to avoid conflicts with existing OMOP vocabularies [15]. For missing source codes, for example, for hierarchy levels, we used the prefix "OMOP" followed by number sequence as the *concept_code*, similar to the *concept_codes* of the OMOP CDM vocabulary OMOP Extension; this procedure follows the convention that a combination of *vocabulary_id* and *concept_code* is supposed to be unique and serves as a secondary key. Furthermore, we had to assign *domains* and *concept_classes* to the new concepts. This was done by searching the English designation of the parent categories of the source codes in Athena and comparing it with the proposed concepts. After a review process consisting of 2 researchers, we decided on a suitable *domain* and *concept_class* and added them to the parent categories as well as to the specific source codes. After the preparation of the *concept* table was completed, we stored the hierarchical structure of the vocabulary, as well as the information about replaced invalid concepts (eg, due to a change

in the designation of a source code), in the *concept_relationship* table.

Evaluation

After the vocabulary preparation, we evaluated the mapping results from Usagi as well as the new 2-billion concepts regarding the completeness of the source codes and the correctness of the validity periods using the synthetic claims data. For both criteria, we implemented an evaluation process with Pentaho Data Integration (Hitachi Vantara) [18]. The input of this process was a list of source codes and the corresponding dates extracted from the synthetic claims data set, as well as the prepared *source_to_concept_map* or *concept* table as CSV files for each vocabulary. Consequently, no connection to an OMOP CDM database was required. Both evaluations were done by an interdisciplinary team in an iterative process until all identified errors had been resolved.

The aim of the completeness assessment was to check if all source codes are available in the prepared vocabularies. This was done by searching the source codes in the *source_code* (*source_to_concept_map*) or *concept_code* (*concept*) columns. If the result of the search had found a *concept_id*, then the codes were already present in the vocabulary; otherwise, they were missing. For both findings, the occurrence of the unique codes in the source record was then calculated and exported as CSV files as a result. In the next step, we further analyzed the missing codes—whether they were forgotten during preparation or if they displayed data quality issues—by using the vocabulary preparation approach for the missing codes, as illustrated in Figure 3.

The second evaluation focused on the correct assignment of the validity periods of the source codes in the prepared vocabularies. For this purpose, we used the source codes for which a matching *concept_id* was found during the first evaluation step. During the correctness assessment, we further checked whether the source codes with their associated dates were valid in the prepared vocabularies. The verification was based on a lookup to identify if the date used in the source was within the validity period of the vocabulary. A positive result indicated that codes were valid, and a negative result indicated that codes were invalid. For both results, we again calculated the occurrence of the unique combination of source code and date in the source and exported the results as CSV files. Afterward, we examined the invalid codes and rechecked whether the validity periods in the vocabularies contained errors.

Final Determination of the Vocabulary Coverage

According to the approach used in Henke et al [12] to calculate the initial vocabulary coverage in the OMOP CDM, we recalculated the vocabulary coverage after the vocabulary preparation. For this purpose, we took our initial list of all German vocabularies used in the claims data set and assigned them again to 3 categories: “available in Athena,” “available through interim mapping,” and “not available.” With regard to the new 2-billion concepts, we added a new fourth category called “Athena-ready.” This category was intended to show that for the new 2-billion concepts, only the external step of loading them into Athena remains. During the final determination, a score of 1 was assigned to a vocabulary if it belongs to the category, and 0 was assigned otherwise. Based on the scores, we calculated the percentage distribution among the 4 categories.

Results

Vocabulary Preparation

Within the selective search, we were able to collect information for all of the missing vocabularies shown in Figure 2 (see Multimedia Appendix 1). However, due to license restrictions, 2 vocabularies (German Diagnosis Related Groups and *Pharmazentralnummer* [Central Pharmaceutical Number]) are legally not allowed to be prepared for the OMOP CDM. For the preparation of the other missing vocabularies, we applied both approaches by using Usagi to map the OMOP CDM standard concepts and by creating new 2-billion concepts. From Table 1, it can be seen that we applied a mapping to standard concepts for 3 vocabularies: *Ambulante spezialfachärztliche Versorgung* (ASV; outpatient specialist care), diagnosis type (inpatient), and provider specialty. These vocabularies were prepared with Usagi when a comprehensive mapping to OMOP CDM standard concepts was possible. The resulting Usagi export of the *source_to_concept_map* table for each vocabulary can be found in our GitHub repository [19].

Table . Overview of vocabularies prepared with Usagi, including number of records in the *source_to_concept_map* table.

Vocabulary name	*Source_to_concept_map*, records, n
ASV[a]	68
Diagnosis type (inpatient)	6
Provider specialty	104

[a]ASV: Ambulante spezialfachärztliche Versorgung (outpatient specialist care).

For the remaining 8 vocabularies, we created new 2-billion concepts to map the source codes to unique *concept_ids* that are analogous to other nonstandard concepts in the OMOP CDM and to prevent different semantic meanings of source codes according to the example shown in Figure 1. Table 2 provides an overview of the number of newly added records for the *vocabulary*, *concept_class*, *concept*, and *concept_relationship* tables for each new 2-billion concept (see the GitHub repository [20] for details). We have prepared almost all vocabularies entirely, that is, all source codes contained in the vocabularies. The exceptions are the *Einheitlicher Bewertungsmaßstab* (EBM; German Uniform Assessment Standard) and inpatient charge types vocabularies. In the case of EBM, we prepared the EBM vocabulary widely used in Germany for the OMOP CDM. For inpatient charge types, we restricted the preparation to the

hierarchy levels “daily charges” and “case-related charges” that were relevant in the Hybrid-QI project.

Table . Overview of vocabularies prepared as new nonstandard concepts, including number of records in the *vocabulary*, *concept_class*, *concept*, and *concept_relationship* tables.

Vocabulary name	*Vocabulary*, records, n	*Concept_class*, records, n	*Concept*, records, n	*Concept_relationship*, records, n
HMK[a]	1	2	263	396
HPNR[b]	1	0	732	1344
PIA[c]	1	0	74	0
EBM[d]	1	2	3614	7348
Inpatient charge types	1	2	992	1997
Outpatient charge types (including EBM)	1	2	5647	11,280
Diagnosis type (outpatient)	0	0	4	0
Diagnosis type (inpatient)	0	0	2	0

[a]HMK: Heilmittelkatalog (catalog of physiotherapy, podological therapy, speech therapy, occupational therapy, and nutritional therapy).

[b]HPNR: Heilmittelpositionsnummern (uniform list of item numbers for therapeutic services).

[c]PIA: Bundeseinheitlicher Katalog für die Dokumentation der Leistungen der psychiatrischen Institutsambulanzen (standardized federal catalog for the documentation of services provided by psychiatric institutional outpatient clinics).

[d]EBM: Einheitlicher Bewertungsmaßstab (German Uniform Assessment Standard).

Evaluation Outcomes

For both variants of the vocabulary preparation through the *source_to_concept_map* or *concept* table, we evaluated completeness and correctness. The evaluation process implemented for this purpose has been released as a GitHub repository [20]. The process includes the extraction of source code and date information from synthetic German claims data, as well as the completeness and correctness assessment for each vocabulary. The following sections demonstrate the final results of the evaluation after multiple iterations.

Completeness

The results of the vocabulary completeness evaluation are summarized in Table 3. From the results, it can be seen that 44.3% (3288/7417) of the source codes can be found in the prepared concepts for the OMOP CDM. What stands out in the table are the EBM and ASV vocabularies. For the EBM vocabulary, 68.21% (4074/5973) of the unique codes were not found in the prepared concepts, that is, they had no *concept_id*, which resulted in 21.68% (1,422,808/6,563,865) of source records not having a *concept_id*. For the ASV vocabulary, no *concept_id* was found for 21.74% (5/23) of the unique codes. Consequently, 8.07% (44/545) of the source records did not have a *concept_id*. Furthermore, we summed up the number of unique codes and total records of the provider specialty vocabulary in Table 3. This is because the vocabulary occurred in 5 different source tables that include data about remedies, outpatient drug prescriptions, and contract medical care.

Table . Results of the vocabulary completeness evaluation.

Vocabulary name	Unique source codes without *concept_id*, n/N[a] (%)	Records without *concept_id*, n/N[b] (%)
HMK[c]	12/116 (10.34)	120/70,514 (0.17)
HPNR[d]	1/105 (0.95)	5/137,106 (0.0036)
PIA[e]	0/37 (0)	0/8721 (0)
EBM[f]	4074/5973 (68.21)	1,422,808/6,563,865 (21.68)
ASV[g]	5/23 (21.74)	44/545 (8.07)
Inpatient charge types	0/101 (0)	0/1222 (0)
Outpatient charge types	37/1062 (3.48)	1760/48,422 (3.63)
Diagnosis type (outpatient)	0/3 (0)	0/17,437 (0)
Diagnosis type (inpatient)	0/6 (0)	0/864,764 (0)
Provider specialty	0/104 (0)	0/9,057,555 (0)

[a]n=number of unique source codes with *concept_id*; N=number of unique source codes.

[b]n=number of records without *concept_id*; N=number of records.

[c]HMK: Heilmittelkatalog (catalog of physiotherapy, podological therapy, speech therapy, occupational therapy, and nutritional therapy).

[d]HPNR: Heilmittelpositionsnummern (uniform list of item numbers for therapeutic services).

[e]PIA: Bundeseinheitlicher Katalog für die Dokumentation der Leistungen der psychiatrischen Institutsambulanzen (standardized federal catalog for the documentation of services provided by psychiatric institutional outpatient clinics).

[f]EBM: Einheitlicher Bewertungsmaßstab (German Uniform Assessment Standard).

[g]ASV: Ambulante spezialfachärztliche Versorgung (outpatient specialist care).

Correctness

The correctness evaluation was performed for the source codes for which a *concept_id* was found during the completeness evaluation. Furthermore, we excluded records if they had a missing source code or date information in the source. For the diagnosis type and provider specialty vocabularies, we were unable to evaluate the correctness due to having no equivalent dates in the source. In addition, these vocabularies have default values for *valid_start_date* and *valid_end_date* in the OMOP CDM. Table 4 shows the results obtained from the correctness evaluation. Looking at the percentage of invalid unique source code–date combinations, for all 7 vocabularies, less than 1% (501/706,032) of the combinations were invalid. In comparison, considering the total number of records, less than 1% (3036/5,358,996) of the records were identified as invalid. The vocabulary with the lowest correctness was HPNR with 0.13% (442/33,439) of invalid unique source code–date combinations and 2.14% (2938/137,101) of invalid records.

Table . Results of the vocabulary correctness evaluation.

Vocabulary Name	Invalid unique source code–date combination, n/N[a] (%)	Invalid records, n/N[b] (%)
HMK[c]	2/27,197 (0.0074)	2/70,394 (0.0028)
HPNR[d]	442/33,439 (0.13)	2938/137,101 (2.14)
PIA[e]	0/2449 (0)	0/8721 (0)
EBM[f]	56/641,285 (0.0087)	95/5,141,057 (0.0018)
ASV[g]	0/490 (0)	0/501(0)
Inpatient charge types	1/1172 (0.09)	1/1222 (0.08)
Outpatient charge types	0/39,123 (0)	0/46,662 (0)

[a]n=number of invalid unique source–date combinations; N=number of unique source–date combinations.

[b]n=number of invalid records; N=number of records.

[c]HMK: Heilmittelkatalog (catalog of physiotherapy, podological therapy, speech therapy, occupational therapy, and nutritional therapy).

[d]HPNR: Heilmittelpositionsnummern (uniform list of item numbers for therapeutic services).

[e]PIA: Bundeseinheitlicher Katalog für die Dokumentation der Leistungen der psychiatrischen Institutsambulanzen (standardized federal catalog for the documentation of services provided by psychiatric institutional outpatient clinics).

[f]EBM: Einheitlicher Bewertungsmaßstab (German Uniform Assessment Standard).

[g]ASV: Ambulante spezialfachärztliche Versorgung (outpatient specialist care).

Vocabulary Coverage

After the evaluation of the prepared vocabularies, we checked the impact of the standard concept mappings and the new 2-billion concepts on the vocabulary coverage in the OMOP CDM. Table 5 shows the results of the final determination of the vocabulary coverage (see Multimedia Appendix 2). As can be seen, the percentage of unavailable vocabularies decreased noticeably from 55% (11/20) to 10% (2/20). One reason for this is the increase of interim mappings from 30% (6/20) to 45% (9/20) using the *source_to_concept_map* table. Another impact has been the creation of new 2-billion concepts for the OMOP CDM. With this approach, 30% (6/20) of the vocabularies were prepared for the OMOP CDM as Athena-ready concepts. The remaining 10% (2/20), which refers to the unavailable vocabularies, is because of the 2 missing vocabularies due to license restrictions.

Table . Vocabulary coverage of German claims data in the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) after vocabulary preparation.

Vocabulary coverage status	Vocabularies (N=20), n (%)
Available in Athena	3 (15)
Athena ready	6 (30)
Available through interim mapping	9 (45)
Not available	2 (10)

Discussion

Vocabulary Preparation Approach

With our presented approach of vocabulary preparation consisting of mapping to standard concepts and the manual creation of new 2-billion concepts, we could improve the vocabulary coverage of missing German vocabularies for use within the OMOP CDM. By preparing 10 vocabularies, we were able to show that our approach is applicable to any type of vocabulary used in a source data set. Referring to our objectives, we succeeded in mapping the majority of source codes used in German claims data to matching *concept_ids* in the OMOP CDM (objective 1). Furthermore, by mapping source codes to standard concepts and creating new 2-billion concepts in the OMOP CDM, we are now able to use German claims data for research based on the OMOP CDM (objective 2).

Nevertheless, there are limitations to our work. The first limitation relates to the reusability of the prepared vocabularies for other researchers. Our newly created 2-billion concepts are not currently available in Athena. Until the integration into Athena is done, the prepared vocabularies are available for other researchers via download from our GitHub repository [19]. When using our vocabularies in the OMOP CDM at other sites, it should be considered that conflicts with 2-billion concepts from other researchers must be avoided by agreeing on certain number ranges within the 2-billion range. A solution for this would be to set up a blank OMOP CDM database, where only the vocabularies provided by Athena are present and our prepared vocabularies are loaded afterward.

Furthermore, our results are currently limited to German vocabularies, which can only be used in national research projects using the OMOP CDM, because the mapping of new

2-billion concepts to standard concepts has not yet taken place. Consequently, the data cannot be used for international studies that are based on OMOP CDM standard concepts. To be able to participate in OHDSI network studies with German claims data, future work is required to map the prepared 2-billion concepts to standard concepts.

Vocabulary Completeness

Our vocabulary completeness evaluation showed that we are currently covering 44.3% (3288/7417) of the codes used in the synthetic German claims data set. Nevertheless, even after multiple iterative adjustments of the vocabularies for missing source codes, some source codes could still not be found in the prepared vocabularies, especially for the Heilmittelkatalog (HMK; catalog of physiotherapy, podological therapy, speech therapy, occupational therapy, and nutritional therapy), HPNR, EBM, ASV, and outpatient charge types vocabularies. The reasons for missing codes can be divided into 2 categories: data quality issues and vocabulary scope limitation. The category of data quality issues summarizes causes such as transposed digits, missing characters, or coding errors (the number zero instead of the letter "O"). Thus, these errors do not refer to incomplete prepared vocabularies but to documentation mistakes in the source. For this reason, the errors are reflected back to the data-providing site, to serve as a basis for future work regarding the development of methods to increase the documentation quality, for example, through a preprocessing of data before they are loaded into the OMOP CDM.

The category of vocabulary scope limitation mainly refers to the EBM vocabulary. For our purpose, we prepared the EBM vocabulary widely used in Germany for the OMOP CDM. However, there also exist local codes administered by the regional associations of statutory health insurance physicians, which can serve as extensions or substitutions to the Germany-wide EBM. The preparation of local EBM codes will be a part of future work.

During the implementation of the evaluation process and the analysis of the results, we encountered challenges for the outpatient charge types vocabulary. The outpatient charge types vocabulary refers to a selection of EBM codes. For this reason, 2 prepared vocabularies had to be considered in the evaluation of completeness. However, for codes that could not be found in either of the 2 vocabularies, it was not possible to assess from which vocabulary they originated. In our next steps of semantic mapping of German claims data to the OMOP CDM, we plan to write such unassignable codes with a *concept_id* of 0 to the OMOP CDM to highlight this issue.

Vocabulary Correctness

The results of the vocabulary correctness evaluation revealed invalid source codes for the HMK, HPNR, EBM, and inpatient charge types vocabularies. Looking at it in more detail, we found that both reasons for invalid codes, that is, invalid codes because of dates before the *valid_start_date* in the OMOP CDM and invalid codes because of dates after the *valid_end_date* in the OMOP CDM, occurred during our evaluation. Although the first problem occurred for all 4 vocabularies, the last 1 only occurred for the EBM vocabulary. A possible explanation for this might be the difference between the date of service provision and the date of payment in the source. Ditscheid et al [21] have already highlighted the temporal discrepancies of these 2 dates and their influence on the interpretation of results. They found that discrepancies in time could lead to an underestimation or overestimation of health service utilization regarding the death date of a patient or the change between years. They proposed to take these discrepancies "into account when requesting the data, but also in preparing and analyzing them" [21]. Consequently, when conducting research with (German) claims data based on the OMOP CDM, we must decide individually for each vocabulary which dates are appropriate for checking their validity in the OMOP CDM concept. However, since only a single piece of date information per vocabulary was available in the synthetic German claims data set, the evaluation conducted in this paper was limited to this information and should be repeated in the future in a second evaluation with more suitable date information per vocabulary.

Comparison With Prior Work

Our proposed vocabulary preparation approach is consistent with methods used by other researchers. There are many papers describing the mapping of source codes or even free texts to OMOP CDM standard concepts using Usagi [8,13,22-25]. However, many papers also describe the approach of creating new 2-billion concepts, since mapping to standard concepts is not always possible [7,9,13,26-31]. For example, Fischer et al [28] created custom concepts for the Pulmonary Hypertension Nice classification. Rinner et al [29] focused on the missing vocabulary of the Austrian pharmaceutical registration number and consequently created new records for the *vocabulary*, *concept_class*, and *concept* tables in the OMOP CDM. Sathappan et al [13] created new unique 2-billion *concept_ids* to store local questionnaire terms in the OMOP CDM. However, none of these approaches describe in detail the process of preparing missing vocabularies for the OMOP CDM. The paper by Sathappan et al [13] laid out a promising approach. However, newly created concepts were directly added as standard concepts to the OMOP CDM, which limits their use to local analysis.

Compared to other research, our approach offers 2 advantages. In terms of a guideline, the presented approach enables the preparation of missing vocabularies for the OMOP CDM for a specific site's data as well as their evaluation. Furthermore, making the vocabularies available via GitHub enables distribution and direct use of the newly created vocabularies for German data and, thus, ensures semantic interoperability across institutions in Germany.

Conclusions

With our presented vocabulary preparation approach, we took a first promising step toward using German claims data for research based on the OMOP CDM. German health care providers, institutes, and health insurance companies can use the prepared German vocabularies, as these vocabularies are part of the legal data transmission for billing processes with health insurance companies. However, the proportion of newly created 2-billion concepts cannot yet be used for international studies due to a missing mapping to standard concepts. During our next steps, we will address this problem by using Usagi to

propose a mapping from the new 2-billion concepts to standard concepts. By doing so, we also want to investigate how well the new German 2-billion concepts can be mapped to OMOP CDM standard concepts and identify the reasons why specific German codes could not be mapped (eg, missing semantic concepts or specifics of the German health care system [billing focus]). In addition, we also aim to collaborate with OHDSI to have our prepared vocabularies externally validated and subsequently integrated into Athena.

Acknowledgments

The research reported in this work was accomplished as part of the German Innovations Fund of the Federal Joint Committee in Germany (G-BA; grant 01VSF20013). The article processing charge was funded by the joint publication funds of the Technische Universität Dresden, including the Carl Gustav Carus Faculty of Medicine, and the Sächsische Landesbibliothek – Staats- und Universitätsbibliothek, Dresden, as well as the Open Access Publication Funding of the Deutsche Forschungsgemeinschaft.

Authors' Contributions

All authors contributed substantially to this work. EH contributed to writing—original draft preparation. EH and LAL contributed to table preparation and translation. EH and MZ reviewed the mapping assumptions. EH contributed to the implementation of the evaluation process. EH, TR, and M Spoden reviewed the evaluation results. EH, MZ, MK, TR, LAL, M Spoden, CG, M Sedlmayr, and FB contributed to writing—review and editing. M Sedlmayr contributed resources. All authors have read and agreed to the current version of the manuscript and take responsibility for the scientific integrity of the work.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Overview of relevant links to the vocabularies considered during the preparation process.
[DOCX File, 15 KB - medinform_v11i1e47959_app1.docx]

Multimedia Appendix 2
Results of the final determination of vocabulary coverage.
[XLSX File, 17 KB - medinform_v11i1e47959_app2.xlsx]

References

1. Office of the Commissioner. Real-world evidence. US Food and Drug Administration. 2023 May 2. URL: www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence [accessed 2023-05-04]
2. Code C03. Athena – OHDSI vocabularies repository. URL: athena.ohdsi.org/search-terms/terms?query=C03 [accessed 2023-03-23]
3. Reinecke I, Zoch M, Reich C, Sedlmayr M, Bathelt F. The usage of OHDSI OMOP - a scoping review. Stud Health Technol Inform 2021 Sep 21;283:95-103. [doi: 10.3233/SHTI210546] [Medline: 34545824]
4. Garza M, del Fiol G, Tenenbaum J, Walden A, Zozus MN. Evaluating common data models for use with a longitudinal community registry. J Biomed Inform 2016 Dec;64:333-341. [doi: 10.1016/j.jbi.2016.10.016] [Medline: 27989817]
5. Athena – OHDSI vocabularies repository. URL: athena.ohdsi.org/ [accessed 2022-11-25]
6. Bardenheuer K, van Speybroeck M, Hague C, Nikai E, Price M. Haematology Outcomes Network in Europe (HONEUR)-a collaborative, interdisciplinary platform to harness the potential of real-world data in hematology. Eur J Haematol 2022 Aug;109(2):138-145. [doi: 10.1111/ejh.13780] [Medline: 35460296]
7. Biedermann P, Ong R, Davydov A, Orlova A, Solovyev P, Sun H, et al. Standardizing registry data to the OMOP Common Data Model: experience from three pulmonary hypertension databases. BMC Med Res Methodol 2021 Nov 2;21(1):238. [doi: 10.1186/s12874-021-01434-3] [Medline: 34727871]
8. Haberson A, Rinner C, Schöberl A, Gall W. Feasibility of mapping Austrian health claims data to the OMOP Common Data Model. J Med Syst 2019 Sep 7;43(10):314. [doi: 10.1007/s10916-019-1436-9] [Medline: 31494719]
9. Lamer A, Abou-Arab O, Bourgeois A, Parrot A, Popoff B, Beuscart JB, et al. Transforming anesthesia data into the Observational Medical Outcomes Partnership Common Data Model: development and usability study. J Med Internet Res 2021 Oct 29;23(10):e29259. [doi: 10.2196/29259] [Medline: 34714250]
10. Spoden M, Dröge P, Roessler M, Datzmann T, Lang C, Sedlmayr M, et al. Hybride qualitätsindikatoren mittels machine learning-methoden (Hybrid-QI). Presented at: 21 Deutscher Kongress für Versorgungsforschung (DKVF 2022); Oct 5-7, 2022; Potsdam, Germany. [doi: 10.3205/22dkvf107]

11. Peng Y, Henke E, Reinecke I, Zoch M, Sedlmayr M, Bathelt F. An ETL-process design for data harmonization to participate in international research with German real-world data based on FHIR and OMOP CDM. Int J Med Inform 2023 Jan;169:104925. [doi: 10.1016/j.ijmedinf.2022.104925] [Medline: 36395615]
12. Henke E, Zoch M, Reinecke I, Spoden M, Ruhnke T, Günster C, et al. German claims data for real-world research: content coverage evaluation in OMOP CDM. Stud Health Technol Inform 2023 May 18;302:3-7. [doi: 10.3233/SHTI230053] [Medline: 37203598]
13. Sathappan SMK, Jeon YS, Dang TK, Lim SC, Shao YM, Tai ES, et al. Transformation of electronic health records and questionnaire data to OMOP CDM: a feasibility study using SG_T2DM dataset. Appl Clin Inform 2021 Aug;12(4):757-767. [doi: 10.1055/s-0041-1732301] [Medline: 34380168]
14. OMOP CDM v5.3 - concept. OMOP Common Data Model. URL: ohdsi.github.io/CommonDataModel/cdm53.html#CONCEPT [accessed 2023-03-23]
15. OMOP CDM frequently asked questions. OMOP Common Data Model. URL: ohdsi.github.io/CommonDataModel/faq.html [accessed 2023-03-13]
16. Schuemie M, Li W, Rijnbeek P, Borgdorff J, Voss E. Usagi. GitHub. 2021 Apr 9. URL: github.com/OHDSI/Usagi [accessed 2023-03-03]
17. Usagi. OHDSI Usagi. URL: ohdsi.github.io/Usagi/ [accessed 2023-10-26]
18. Pentaho data integration. Hitachi Vantara. 2022 Nov 17. URL: help.hitachivantara.com/Documentation/Pentaho/9.4/Products/Pentaho_Data_Integration [accessed 2023-03-03]
19. Henke E. OMOP-CDM-German-vocabularies. GitHub. 2023 Mar 28. URL: github.com/elisahenke/OMOP-CDM-German-vocabularies [accessed 2023-03-24]
20. Henke E. OMOP-vocabulary-evaluation. GitHub. 2023 Mar 24. URL: github.com/elisahenke/OMOP-vocabulary-evaluation [accessed 2023-03-24]
21. Ditscheid B, Storch J, Krause M, Meyer I, Freytag A. Leistungs- und abrechnungsdatum in GKV-routinedaten: umgang mit zeitlichen abweichungen. Date of service provision and date of payment in claims data: dealing with time differences. Artice in German. Gesundheitswesen 2020 Mar;82(S 01):S20-S28. [doi: 10.1055/a-1030-4223] [Medline: 31822022]
22. Papez V, Moinat M, Payralbe S, Asselbergs FW, Lumbers RT, Hemingway H, et al. Transforming and evaluating electronic health record disease phenotyping algorithms using the OMOP Common Data Model: a case study in heart failure. JAMIA Open 2021 Feb 4;4(3):ooab001. [doi: 10.1093/jamiaopen/ooab001] [Medline: 34514354]
23. Papez V, Moinat M, Voss EA, Bazakou S, van Winzum A, Peviani A, et al. Transforming and evaluating the UK Biobank to the OMOP Common Data Model for COVID-19 research and beyond. J Am Med Inform Assoc 2022 Dec 13;30(1):103-111. [doi: 10.1093/jamia/ocac203] [Medline: 36227072]
24. Almeida JR, Silva JF, Matos S, Oliveira JL. A two-stage workflow to extract and harmonize drug mentions from clinical notes into observational databases. J Biomed Inform 2021 Aug;120:103849. [doi: 10.1016/j.jbi.2021.103849] [Medline: 34214696]
25. Blacketer C, Voss EA, DeFalco F, Hughes N, Schuemie MJ, Moinat M, et al. Using the Data Quality Dashboard to improve the EHDEN network. Applied Sciences 2021 Dec 15;11(24):11920. [doi: 10.3390/app112411920]
26. Lima DM, Rodrigues JFJ, Traina AJM, Pires FA, Gutierrez MA. Transforming two decades of ePR data to OMOP CDM for clinical research. Stud Health Technol Inform 2019 Aug 21;264:233-237. [doi: 10.3233/SHTI190218] [Medline: 31437920]
27. Kim JW, Kim S, Ryu B, Song W, Lee HY, Yoo S. Transforming electronic health record polysomnographic data into the Observational Medical Outcome Partnership's Common Data Model: a pilot feasibility study. Sci Rep 2021 Mar 29;11(1):7013. [doi: 10.1038/s41598-021-86564-w] [Medline: 33782494]
28. Fischer P, Stöhr MR, Gall H, Michel-Backofen A, Majeed RW. Data integration into OMOP CDM for heterogeneous clinical data collections via HL7 FHIR bundles and XSLT. Stud Health Technol Inform 2020 Jun 16;270:138-142. [doi: 10.3233/SHTI200138] [Medline: 32570362]
29. Rinner C, Gezgin D, Wendl C, Gall W. A clinical data warehouse based on OMOP and i2b2 for Austrian health claims data. Stud Health Technol Inform 2018;248:94-99. [doi: 10.3233/978-1-61499-858-7-94] [Medline: 29726424]
30. Lamer A, Depas N, Doutreligne M, Parrot A, Verloop D, Defebvre MM, et al. Transforming French electronic health records into the Observational Medical Outcome Partnership's Common Data Model: a feasibility study. Appl Clin Inform 2020 Jan;11(1):13-22. [doi: 10.1055/s-0039-3402754] [Medline: 31914471]
31. Paris N, Lamer A, Parrot A. Transformation and evaluation of the MIMIC database in the OMOP Common Data Model: development and usability study. JMIR Med Inform 2021 Dec 14;9(12):e30970. [doi: 10.2196/30970] [Medline: 34904958]

Abbreviations

ASV: Ambulante spezialfachärztliche Versorgung (outpatient specialist care)
CDM: common data model
EBM: Einheitlicher Bewertungsmaßstab (German Uniform Assessment Standard)

HMK: Heilmittelkatalog (catalog of physiotherapy, podological therapy, speech therapy, occupational therapy, and nutritional therapy)
HPNR: Heilmittelpositionsnummern (uniform list of item numbers for therapeutic services)
Hybrid-QI: Hybrid Quality Indicators Using Machine Learning Methods
***ICD-10-GM*:** *International Classification of Diseases, Tenth Revision, German Edition*
OHDSI: Observational Health Data Sciences and Informatics
OMOP: Observational Medical Outcomes Partnership
RWD: real-world data

Edited by C Lovis, C Perrin; submitted 06.04.23; peer-reviewed by C Reich, S Sarbadhikari; revised version received 07.09.23; accepted 09.09.23; published 07.11.23.

Please cite as:
Henke E, Zoch M, Kallfelz M, Ruhnke T, Leutner LA, Spoden M, Günster C, Sedlmayr M, Bathelt F
Assessing the Use of German Claims Data Vocabularies for Research in the Observational Medical Outcomes Partnership Common Data Model: Development and Evaluation Study
JMIR Med Inform 2023;11:e47959
URL: https://medinform.jmir.org/2023/1/e47959
doi:10.2196/47959

Original Paper

A Standardized Clinical Data Harmonization Pipeline for Scalable AI Application Deployment (FHIR-DHP): Validation and Usability Study

Elena Williams[1], MSc; Manuel Kienast[1], MSc; Evelyn Medawar[1], MSc; Janis Reinelt[1], MD; Alberto Merola[1], PhD; Sophie Anne Ines Klopfenstein[2], MD; Anne Rike Flint[2], MSc; Patrick Heeren[2], BSc; Akira-Sebastian Poncette[2], MD, PD; Felix Balzer[2], Prof Dr; Julian Beimes[3], MSc; Paul von Bünau[3], PhD; Jonas Chromik[4], MSc; Bert Arnrich[4], Prof Dr; Nico Scherf[5], PhD; Sebastian Niehaus[1], MSc

[1]AICURA Medical GmbH, Berlin, Germany

[2]Institute of Medical Informatics, Charité – Universitätsmedizin Berlin, Berlin, Germany

[3]idalab GmbH, Berlin, Germany

[4]Digital Health – Connected Healthcare, Hasso Plattner Institute, University of Potsdam, Potsdam, Germany

[5]Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany

Corresponding Author:
Evelyn Medawar, MSc
AICURA Medical GmbH
Bessemerstr 22
Berlin, 12103
Germany
Phone: 49 173 9449677
Email: evelyn.medawar@aicura-medical.com

Abstract

Background: Increasing digitalization in the medical domain gives rise to large amounts of health care data, which has the potential to expand clinical knowledge and transform patient care if leveraged through artificial intelligence (AI). Yet, big data and AI oftentimes cannot unlock their full potential at scale, owing to nonstandardized data formats, lack of technical and semantic data interoperability, and limited cooperation between stakeholders in the health care system. Despite the existence of standardized data formats for the medical domain, such as Fast Healthcare Interoperability Resources (FHIR), their prevalence and usability for AI remain limited.

Objective: In this paper, we developed a data harmonization pipeline (DHP) for clinical data sets relying on the common FHIR data standard.

Methods: We validated the performance and usability of our FHIR-DHP with data from the Medical Information Mart for Intensive Care IV database.

Results: We present the FHIR-DHP workflow in respect of the transformation of “raw” hospital records into a harmonized, AI-friendly data representation. The pipeline consists of the following 5 key preprocessing steps: querying of data from hospital database, FHIR mapping, syntactic validation, transfer of harmonized data into the patient-model database, and export of data in an AI-friendly format for further medical applications. A detailed example of FHIR-DHP execution was presented for clinical diagnoses records.

Conclusions: Our approach enables the scalable and needs-driven data modeling of large and heterogenous clinical data sets. The FHIR-DHP is a pivotal step toward increasing cooperation, interoperability, and quality of patient care in the clinical routine and for medical research.

(JMIR Med Inform 2023;11:e43847) doi:10.2196/43847

KEYWORDS

data interoperability; fast healthcare interoperability resources; FHIR; data standardization pipeline; medical information mart for intensive care; MIMIC IV; artificial intelligence; AI application; AI; deployment; data; usability; care unit; diagnosis; cooperation; patient care; care; medical research

Introduction

The increasing digitalization of health care creates vast amounts of clinical data that are collected and stored in an Electronic Health Record (EHR). Patient information from all medical domains is captured in diverse sets of data recorded in stand-alone systems. With the prevalent use of EHRs in health care organizations, there is abundant opportunity for the additional application of EHR data in clinical and translational research. For instance, such data can be used to develop artificial intelligence (AI) algorithms, which have the potential to transform patient care and medical research. Resource-intensive and inefficient clinical workflows could be optimized by the analysis of historical data with AI applications [1,2]. In particular, the time-consuming and financially costly process of identifying and enrolling the right patients into a clinical trial manually can be reduced significantly by automation [3,4]. However, the exchange of medical data remains limited due to the lack of data interoperability between health care providers, owing to outdated IT infrastructure, inconsistencies in data formats, poor data quality, inadequate data exchange solutions, and data silos [5,6]. To achieve data interoperability, the following steps must be incorporated: (1) integration of isolated data silos, (2) safe exchange of data, and (3) effective use of the available data [7]. Each of these operations includes database schema matching [8] and schema mapping [9], which allow translation of the relationships between the source database and the target data standard.

Employing a harmonized data format will facilitate the exchange of medical data, enabling wide-ranging data-driven collaborations within the private and public health care sectors. Data interoperability requires EHR data to be structured in a common format and in standardized terminologies. Standardization is often performed by adopting the Health Level 7 Fast Healthcare Interoperability Resources (FHIR) model [10], which is supported by numerous health care institutions and vendors of clinical information systems [11]. FHIR is an international industry standard that integrates diverse sets of data in well-defined exchangeable segments of information, which are known as FHIR resources. Therefore, FHIR facilitates interoperability between health care organizations and allows third-party developers to provide medical applications that can be easily integrated into the existing systems. FHIR enables the harmonization of data and thus allows standardized data processing as well as the rollout of AI applications across different clinics and hospitals regardless of which information system they use. Consequently, FHIR forms an important component for the scalable development and deployment of AI in clinics and hospitals.

However, to apply AI, the input data need to be adapted to the AI algorithms. The conventional AI frameworks such as Tensorflow [12] and Pytorch [13] require data to take a tensor form, which is a vector or matrix of n-dimensions that represents various types of data (eg, tabular, time series, image, and text). Since the FHIR format has a multilayered nested structure, a use case–specific data preprocessing is needed. For instance, depending on the AI application and the chosen data source, a custom data preprocessing pipeline should be designed leading to diminished AI scalability. Prior research addressed this problem in different forms but focused on individual applications, thereby constraining the purpose of FHIR to be applicable regardless of the use case [11]. There have been a few attempts to flatten the hierarchical FHIR structure and transform it into NDJSON-based data format [14] or tabular format saved in CSV files [15]. Such formats are more AI-friendly as they represent the data in a more accessible and standardized form for an application of common AI frameworks. Nonetheless, the NDJSON-based FHIR data transformation approach [14] does not provide data selection criteria and filtering capabilities [16]. The approach presented in [15] requires expert knowledge of FHIRPath query language. Moreover, FHIR-based data preprocessing pipelines have been implemented in different contexts, for instance, as electronic data capture [17], as a natural language processing tool [12], and as a standardization protocol based on the Resource Description Framework [6]. Despite the immense benefit they offer regarding processing EHR data, existing approaches are limited to specific use cases or require considerable data preparation to perform standardization. Furthermore, their final output is not easily accessible by common data preprocessing tools and thus hinders the application of AI.

In this paper, we address the challenge of data interoperability in the health care sector by proposing an FHIR data harmonization pipeline (DHP) that provides EHR data in an AI-friendly format. The newly developed FHIR-DHP represents a data workflow solution that includes the aforementioned operations, such as data exchange, mapping, and export. Data privacy is a delicate topic in health care and is of great ethical concern [18]. Given the degree of automation, FHIR-DHP should allow the preprocessing of unseen data in an isolated hospital environment, which makes harmonization privacy preserving.

Methods

Ethical Considerations

The authors did not seek an ethics review board assessment due to the methodology of the study, which included open datasets and data preprocessing pipelines only.

FHIR-DHP Architecture Development

In our work, we propose a generic solution to harmonize hospital EHR data. The FHIR-DHP was designed based on the extract-transform-load framework [19], in which the data are pulled out (ie, queried) from diverse sources, processed into the desired format, and loaded into a data warehouse, namely the

”patient-model” database (DB). As the hospital database contains highly sensitive patient data, it is located behind the hospital’s security infrastructure and is completely isolated from outside access. Therefore, an edge-computation solution was designed, bringing the FHIR-DHP into the hospital’s own infrastructure. The edge-computation solution represents a set of frameworks that perform data querying, preprocessing, storage, and export. In this setting, direct access to the sensitive data is not required to run the standardization pipeline. The queries to the data are defined beforehand based on the database documentation.

To bring the data into a harmonized form, we used an FHIR data model, which is applied by mapping the relationships between the source database and the desired data standard. The FHIR standard is straightforward to implement because it provides a choice of JSON, XML, or resource description format for data representation. The mapping pipeline was developed in the Python programming language to translate queried hospital data into matching FHIR concepts and save the resulting resources in JSON format. The semantics of features from the source database and FHIR concepts are examined using available database and FHIR documentation. The conversion to FHIR was designed to only support a core release 4 standard of the FHIR format to allow generic data preprocessing.

To prevent errors in the remote data standardization scenario, the syntactic validation of FHIR resources is necessary. For instance, the conversion of data types can sometimes lead to erroneous values, especially with date features. Automatic syntactic validation allows the logging of occurred errors and the improvement of harmonization pipeline when working with unseen data. When syntactic validation is completed, FHIR resources should be transferred to the data warehouse to allow the fast and easy retrieval of standardized data for AI applications.

In the final stage of data export, we designed the output that provides the benefits of the original FHIR format with a high level of clinical detail that is also easily accessible for computational tools. We wanted to restructure the data representation in a way that supports effortless data selection and filtering capabilities and would not require a knowledge of FHIRPath query language. Consequently, this output format would enable the smooth conversion of data into a “tensor” format required by conventional AI frameworks.

FHIR-DHP Validation

To demonstrate and evaluate how the FHIR-DHP works, we used the openly available Medical Information Mart for Intensive Care IV (MIMIC IV) database [20]. MIMIC IV includes patient data from the intensive care units at a tertiary academic medical center in Boston, MA, United States. We selected a wide range of tables from MIMIC IV, which cover most of the events occurring during the hospital stay as well as core patient details, information about admissions, and hospital transfers (further referred to as core tables). The event tables include laboratory results, diagnoses, prescriptions, and other details, as shown in Table 1. In addition, MIMIC IV includes the so-called reference tables containing matching dictionaries with medical terms that are used in the hospital records.

Table 1. Selected core and event Medical Information Mart for Intensive Care IV (MIMIC IV) tables as well as the reference dictionary tables that were merged together with core and event tables for Fast Healthcare Interoperability Resources mapping.

Selected core and event MIMIC IV tables	Selected MIMIC IV reference tables
Patient	—[a]
Admissions	—
Transfers	—
Chartevents	d_items
Labevents	d_labitems
Procedureevents	d_items
Prescriptions	—
Inputevents	d_items
Microbiologyevents	—
Outputevents	d_items
Procedures_icd	d_icd_procedures
Diagnoses_icd	d_icd_diagnoses

[a]Not available.

The selected tables were mapped to FHIR standard. Automatic semantic validation is unfeasible, so 2 of the authors manually validated the mapping semantics independently of each other. There are many tools that perform automatic syntactic validation, such as the Python-based package “fhir.resources” used herein [21]. To evaluate the exporting of data from the patient-model DB, we retrieved the diagnosis records.

Results

FHIR-DHP Architecture

The approach presented here represents a scalable protocol for harmonizing hospital EHR data sets based on 5 stages from data query to data export in a standardized format.

Querying Data From the Hospital Database

To connect the FHIR-DHP pipeline to the hospital DB, a communication server is employed. This server runs all necessary queries to retrieve the patient data. The query execution can be run at regular intervals as well as in batches of patients, so as not to overload the data pipeline. Furthermore, the queries prestructure the data according to their semantic relations before proceeding to data mapping.

Mapping Data to FHIR

FHIR allows describing data formats and elements that are recorded as "resources" and an application programming interface for exchanging EHRs. To perform the mappings, semantics of features from the source database and FHIR concepts are explored as well as the relationships between the data tables. Consequently, the mappings between the database tables and FHIR resources are defined. Features where a matching FHIR concept is not found are excluded. The resulting FHIR resources are then saved in JSON format.

Syntactic Validation of FHIR Mappings

During validation, mapped data are ensured to have the correct data types as well as the syntactic format where the hierarchy is maintained, and entries follow FHIR standard specifications. All mappings are validated first during the development stage to identify structural errors and data type inconsistencies. A validation algorithm is incorporated into the pipeline to confirm the correctness of the transformed data in the remote data standardization scenario.

Transferring FHIR Resources to Patient-Model DB

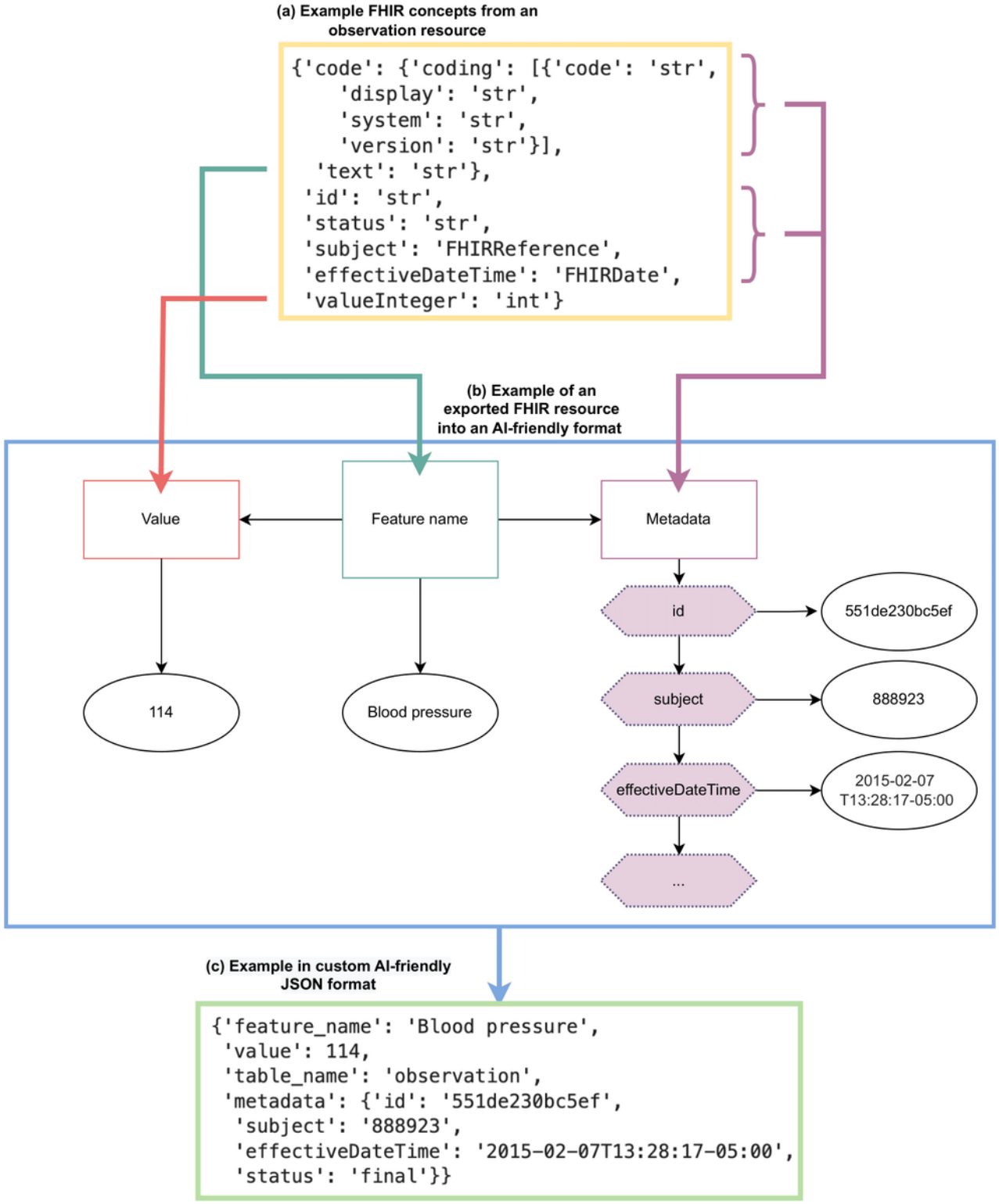
The DB of choice for the patient model is Postgres [22], which is an open-source relational DB management system featuring SQL compliance and storage of JSON documents. The database for the FHIR resources is used to harmonize the locally available data only once to allow the further application of various medical AI-based solutions. The data are stored according to FHIR resource type where each resource is saved in a separate JSON structure.

Exporting Data Into Custom JSON format

To export the data from the patient-model DB, the selection is performed by outlining the tables and features of interest in a configuration file, which is then used to determine which harmonized data should be queried. FHIRPath queries were written to retrieve all elements from FHIR resources adhering to specific formatting rules in respect of the predefined key-value structure and to place the extracted elements into the custom JSON file. Such transformation flattens the hierarchical structure of FHIR resources and makes the data more accessible for common data preprocessing tools. The final flattened output does not require expert knowledge of FHIRPath query language and supports effortless data selection and filtering. The resulting file also allows the uncomplicated conversion of data into a "tensor" format required by conventional AI frameworks and fast data selection based on the following 4 keys: feature_name, table_name, value, and metadata.

In Figure 1, we demonstrate how the FHIR-DHP recodes nested FHIR syntax to more accessible features in an AI-friendly format. Example FHIR concepts from an observation resource are given in Figure 1a, where the code's entity "text" defines the record or measurement label. The entity "text" is often duplicated in the item "display." However, depending on the coding system, this "display" item can change, whereas "text" always stays the same and is therefore used as a feature name. The information from the FHIR resource is grouped into the 4 concept keys of feature name (eg, "Blood pressure"), value (eg, "114"), table name (eg, "observation"), and metadata (Figure 1b). For a given FHIR resource type, the metadata may include concepts such as dates, references, coding system details, and resource ID, among other things. As an output, feature names together with a corresponding value and available metadata are provided in a custom JSON structure (Figure 1c). The defined format allows uncomplicated data selection and aggregation based on resource type (eg, "table_name"), feature name, and value. Additional information in a standardized format can be easily accessed from the metadata key and allows further data manipulation.

Figure 1. Conceptual overview for an exemplary Fast Healthcare Interoperability Resources (FHIR) structure and hospital record, which are transformed from FHIR standard to an artificial intelligence (AI)–friendly format.



FHIR-DHP Validation

The MIMIC IV data were queried accordingly to the defined FHIR mappings. The core and event MIMIC IV tables were merged with reference tables to contain a complete description of the hospital records. As a result, the data were grouped and restructured into the information blocks required in FHIR standard. Manual independent validation of the mapping semantics resulted in slight discrepancies, which were subsequently resolved to adhere closely to the FHIR standard. The automatic syntactic validation allowed the prompt verification of standardization operations.

Table 2 shows to which FHIR resources the MIMIC IV tables were mapped. The largest proportion of tables (4 out of 12 tables) were mapped to the *Observation* FHIR resource type, which included lab, microbiology, output, and charted events

collected throughout the patient's stay. The information on admissions and transfers was translated into the *Encounter* FHIR resource (2 out of 12 tables). Procedure events and International Classification of Diseases codes (2 out of 12 tables) were stored in the *Procedure* FHIR resource. Given that the prescriptions table contains medication requests (1 out of 12 tables) and the input events table holds records of medication administration (1 out of 12 tables), these tables were mapped to the corresponding FHIR resource types. Finally, the *Condition* FHIR resource was used to map the table with the patients' diagnosis details (1 out of 12 tables).

In Table 3, we demonstrate how the mapping of the MIMIC IV "diagnoses_icd" table to *Condition* FHIR resource was conducted. Multiple columns of the "diagnoses_icd" table such as "icd_code", "icd_version," and "long_title" were mapped to the FHIR "condition.code" concept, which has a nested structure and provides keys to store the exact International Classification of Diseases code, the version of the coding system, and the code title. The full diagnosis title was mapped both to the "display" and "text" entities.

Figure 2 shows an example of how queried diagnoses records are harmonized to an AI-friendly format. The standardization follows the FHIR-DHP stages described above. At first, the raw data from tables "diagnoses_icd" and "d_icd_diagnoses" are queried (Figure 2a) and merged accordingly to the defined FHIR mappings. Then, the features are renamed as defined in Table 3 for the FHIR condition resource, and the required entities such as "resourceType" and "id" are created (Figure 2b). Finally, the values are placed into a nested FHIR structure (Figure 2c), and subsequently, the data are transformed into a JSON format (Figure 2d), which can be automatically validated (Figure 2e) and saved in the patient-model DB. When the resource is not approved in terms of its syntactic quality (eg, data type, nested structure, or cardinality), an error is raised, which prevents the further saving of this resource in the patient-model DB (Figure 2e). Otherwise, the resource is transferred into a storage (Figure 2f), and the requested data are exported in a custom AI-friendly JSON format (Figure 2g).

We provide an example of a further 2-step transformation of harmonized diagnosis data to a "tensor" format in Multimedia Appendix 1 [12,23].

Table 2. Overview of the mappings performed on the selected Medical Information Mart for Intensive Care (MIMIC) database (DB) tables to Fast Healthcare Interoperability Resources (FHIR) types.

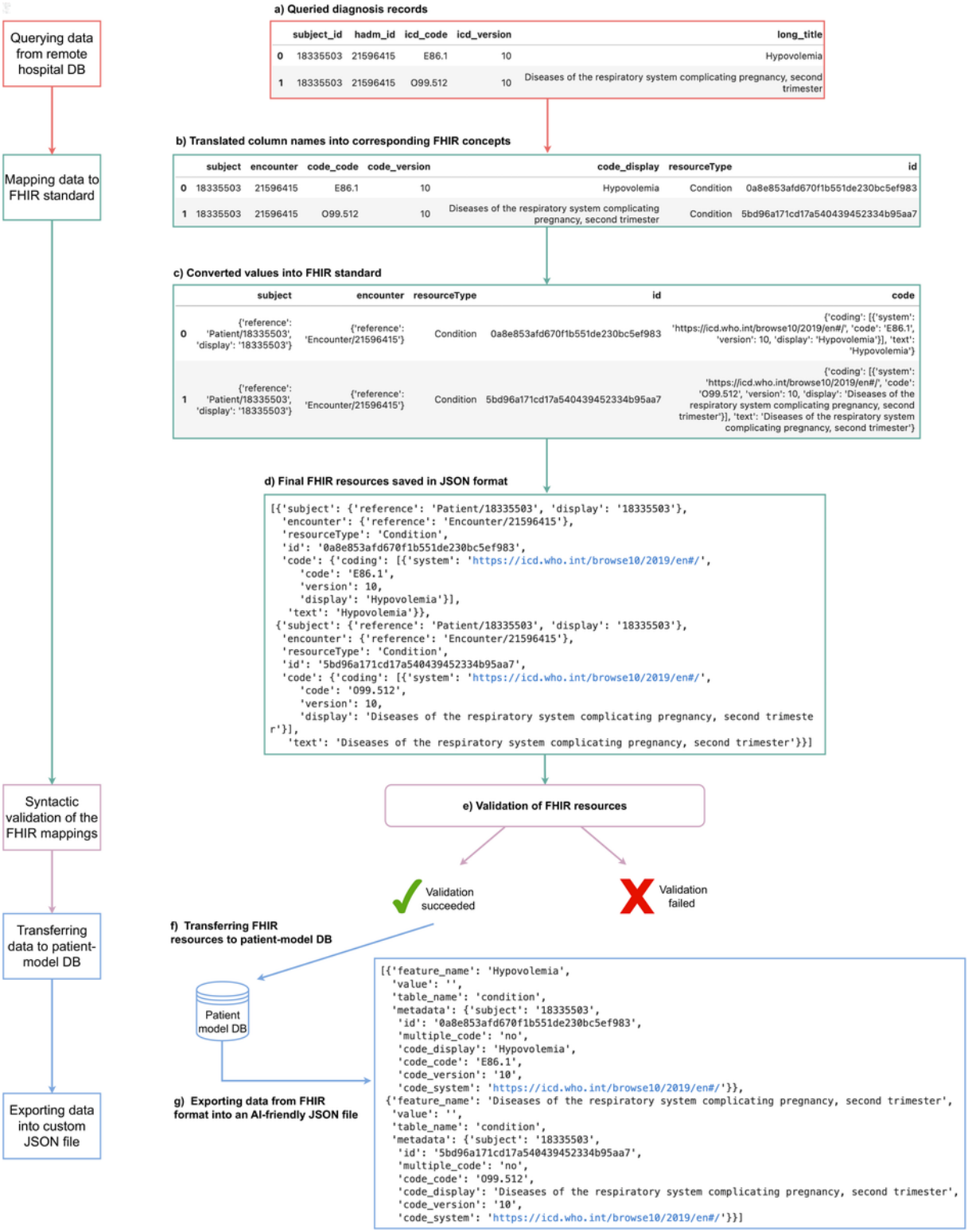
MIMIC IV DB	FHIR resource type
Patients	Patient
Admissions	Encounter
Transfers	Encounter
Chartevents	Observation
Labevents	Observation
Procedureevents	Procedure
Prescriptions	MedicationRequest
Inputevents	MedicationAdministration
Microbiologyevents	Observation
Outputevents	Observation
Procedure_icd	Procedure
Diagnoses_icd	Condition

Table 3. Mapping of "diagnoses_icd" table to Condition Fast Healthcare Interoperability Resources (FHIR) resource.

MIMIC[a] format	FHIR resource format
mimic.diagnoses_icd.subject_id	fhir.condition.subject
mimic.diagnoses_icd.hadm_id	fhir.condition.encounter
mimic.diagnoses_icd.icd_code	fhir.condition.code_code
mimic.diagnoses_icd.icd_version	fhir.condition.code_version
mimic.diagnoses_icd.long_title	fhir.condition.code_display
mimic.diagnoses_icd.long_title	fhir.condition.code_text

[a]MIMIC: Medical Information Mart for Intensive Care.

Figure 2. Flowchart showing an example diagnosis data being processed through the 5 stages in Fast Healthcare Interoperability Resources (FHIR) data harmonization pipeline (DHP). The first stage (a) includes querying of the diagnoses records, at the second stage (b-c) the data are mapped to FHIR standard, and the third stage carries out the syntactic resource validation. (f) If the FHIR resource is successfully validated, it is being transferred into the patient-model database (DB), and then (g) exported in a custom artificial intelligence (AI)–friendly JSON format.



Discussion

Principal Findings

The Harmonization of EHR data is a crucial step toward increasing cooperation, interoperability, and quality of patient care in the clinical routine and medical research. To drive the harmonization of medical data forward, we developed the FHIR-DHP and evaluated it on key MIMIC IV tables. A detailed example of data standardization was presented for clinical diagnosis records from the MIMIC IV database. The FHIR-DHP allows the querying of health data in an isolated environment by employing an edge-computation solution and a communication server, which retrieve patient data and

prestructure it for further mapping to the FHIR standard. A validation step ensures syntactic compliance and initiates the transfer of formatted data to the patient-model DB. The data export provides FHIR resources in a custom JSON file format.

Owing to the FHIR format's multilayered nested structure, its accessibility for AI algorithms is low as it requires transformation into a format compatible with common data preprocessing tools. Thus far, a number of studies have attempted to solve this problem. However, the final output of these studies has not supported data selection criteria and filtering capabilities [14] and requires expert knowledge of FHIRPath query language [15]. In this study, we introduce a custom JSON format that represents a higher level of abstraction to support easier data selection based on the following 4 keys: feature_name, table_name, value, and metadata. Moreover, the newly developed JSON structure fits the expected data format of common data preprocessing frameworks, which are designed to work efficiently with tabular data. As a result, the output presented facilitates the generic and fast deployment of AI and patient cohort identification algorithms.

In comparison to [17,24], the details of FHIR-DHP execution inside the hospital environment in respect of protecting data privacy are discussed. This step, though crucial, is often omitted and left out of the published standardization protocols. The edge-computation solution sets up the FHIR-DHP in a privacy-preserving way where the preprocessing of the patient-related data is performed inside the hospital and is completely isolated from outside access. The so-called federated learning (FL) framework [25] can be integrated into the FHIR-DHP workflow to run algorithms locally, using data from the on-premises database in the respective hospitals and to merge model parameters centrally in the cloud without any patient data leaving the hospital. The FL framework requires data to be in a consistent format across various hospital systems. The developed pipeline achieves such a format and enables the scaling of AI applications.

Thus far, there are only 2 studies attempting to perform the mapping of an MIMIC IV database [26,27]. In [26], the mapping was performed on fewer tables than our approach (8 versus 12 tables). The FHIR mappings from [27] have been recently released and were not yet widely validated. Similar to the approach taken in [17,24,26], FHIR-DHP includes the verification of the performed FHIR mapping, which is essential to ensure the validity of data transformation and to adhere to FHIR version updates. Moreover, in comparison to [17,24,26], FHIR-DHP represents a generic approach to standardize EHR data and can be applied to various hospital database systems.

With the introduction of the FHIR-DHP into the hospital environment, a number of patient-stay parameters can be potentially optimized using AI-based algorithms. For example, the length of stay as well as mortality could be reduced [28], and patients suitable for trial treatment could be automatically and efficiently identified [29]. In consequence, the financial impact on medical providers in respect of personnel time and resources would decrease considerably. The FHIR-DHP aims to bring health care closer to digital transformation and thus toward "Healthcare 4.0" [30] by making EHR data usable "from bedside-to-bench." By inverting the idea of translational research, in contrast to "from bench-to-bedside," accessing the full potential of medical big data with AI will further inform and advance basic research.

Limitations

There are several limitations that we would like to emphasize. FHIR-DHP only works with a core standard of the FHIR format. Those core FHIR resource types have a bounded set of concepts that present a constraint to mapping accuracy. Although the standard resources can be expanded using a profiling technique or FHIR extensions, the use of those would make the FHIR-DHP less generic. Hence, we implemented the mapping using only the standard FHIR resources and omitted some of the MIMIC IV data features that did not have a matching concept in FHIR. Additionally, the FHIR mapping step is subject to the extent of the detail of the database documentation used to infer the semantic and syntactic properties of the data. A solution for an automatic concept recognition can potentially solve this problem. The existing approach in [6] is limited to a small number of FHIR resources and requires an extensive data preparation. Further experiments in this direction could alleviate the concept-matching problem and the requirement for a detailed database description. Moreover, the validation and robustness of FHIR-DHP needs to be tested on other EHR data sets to evaluate its generic setup. In addition, to validate the FHIR-DHP compatibility with machine learning pipelines, further experiments are needed.

Future Prospects

The proposed FHIR-DHP pipeline highlights the therein featured essential data standardization stages and holds the potential to becoming an interoperable harmonization system with an AI-friendly data format. FHIR-DHP enables interoperability and cooperation between clinical institutions and a rapid patient cohort identification for clinical trials; it also unlocks the potential of big medical data.

Conclusions

We provide a comprehensive approach to transforming unstandardized EHR data into a harmonized multilayered nested FHIR format and then to a more readable and more efficient AI-friendly JSON structure. We developed a 5-stage data harmonization pipeline, which includes validation checks. The AI-friendly format of hospital data allows the generic and fast integration of both AI and patient cohort identification algorithms. Harmonized and standardized health care data are of great value to advancing efficiency in big data processing, cooperation, and multicenter data exchange in the clinical sector, boosting medical research, patient care, and clinical trial cohort identification. The next steps would include validating our approach in a hospital environment and applying a privacy-preserving FL framework to make use of advanced AI deployment.

Acknowledgments

This work was partially funded by the German Federal Ministry of Education and Research under Grant 16SV8559.

Availability of Data and Materials

The MIMIC IV database used in this study is openly available to credentialed users who sign the “Data Use Agreement” at PhysioNet website [20]. The code is not publicly available due to privacy, but a demo is available from the corresponding author on request.

Authors' Contributions

EW, SN, MK, JR, and AM were responsible for the study conception; EW and MK took part in data analysis; EW, SN, and EM created the figures; EW, MK, AM, and SN were responsible for methods. EW, EM, JR, and SAIK wrote the draft; BA, JB, PVB, JC, ARF, ASP, and NS reviewed and revised the work.

Conflicts of Interest

FB reports grants from the German Federal Ministry of Education and Research, German Federal Ministry of Health, Berlin Institute of Health, personal fees from Elsevier Publishing, grants from Hans Böckler Foundation, other from Robert Koch Institute, grants from Einstein Foundation, grants from Berlin University Alliance, personal fees from Medtronic and personal fees from GE Healthcare.

Multimedia Appendix 1
Transformation of data saved in custom JSON to tensor format.
[DOCX File , 998 KB - medinform_v11i1e43847_app1.docx]

References

1. Au-Yeung WM, Sahani AK, Isselbacher EM, Armoundas AA. Reduction of false alarms in the intensive care unit using an optimized machine learning based approach. NPJ Digit Med 2019;2:86 [FREE Full text] [doi: 10.1038/s41746-019-0160-7] [Medline: 31508497]
2. Desautels T, Calvert J, Hoffman J, Jay M, Kerem Y, Shieh L, et al. Prediction of Sepsis in the Intensive Care Unit With Minimal Electronic Health Record Data: A Machine Learning Approach. JMIR Med Inform 2016 Sep 30;4(3):e28 [FREE Full text] [doi: 10.2196/medinform.5909] [Medline: 27694098]
3. Maier C, Kapsner LA, Mate S, Prokosch H, Kraus S. Patient Cohort Identification on Time Series Data Using the OMOP Common Data Model. Appl Clin Inform 2021 Jan 27;12(1):57-64 [FREE Full text] [doi: 10.1055/s-0040-1721481] [Medline: 33506478]
4. Ni Y, Wright J, Perentesis J, Lingren T, Deleger L, Kaiser M, et al. Increasing the efficiency of trial-patient matching: automated clinical trial eligibility pre-screening for pediatric oncology patients. BMC Med Inform Decis Mak 2015 Apr 14;15(1):28 [FREE Full text] [doi: 10.1186/s12911-015-0149-3] [Medline: 25881112]
5. de Mello BH, Rigo SJ, da Costa CA, da Rosa Righi R, Donida B, Bez MR, et al. Semantic interoperability in health records standards: a systematic literature review. Health Technol (Berl) 2022;12(2):255-272 [FREE Full text] [doi: 10.1007/s12553-022-00639-w] [Medline: 35103230]
6. Kiourtis A, Mavrogiorgou A, Menychtas A, Maglogiannis I, Kyriazis D. Structurally Mapping Healthcare Data to HL7 FHIR through Ontology Alignment. J Med Syst 2019 Feb 05;43(3):62. [doi: 10.1007/s10916-019-1183-y] [Medline: 30721349]
7. Pagano P, Candela L, Castelli D. Data Interoperability. Data Sci. J 2013;12:GRDI19-GRDI25. [doi: 10.2481/dsj.GRDI-004]
8. Rahm E, Bernstein P. A Survey of Approaches to Automatic Schema Matching. VLDB J 2001;10:334-350. [doi: 10.1007/s007780100057]
9. Kolaitischema M, data E, metadata M. Schema mappings, data exchange, and metadata management. 2005 Presented at: Proceedings of the ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems; June 13-15, 2005; Baltimore, USA p. 61-76. [doi: 10.1145/1065167.1065176]
10. Welcome to FHIR. HL7.org. URL: https://www.hl7.org/fhir/ [accessed 2023-02-15]
11. Vorisek CN, Lehne M, Klopfenstein SAI, Mayer PJ, Bartschke A, Haese T, et al. Fast Healthcare Interoperability Resources (FHIR) for Interoperability in Health Research: Systematic Review. JMIR Med Inform 2022 Jul 19;10(7):e35724 [FREE Full text] [doi: 10.2196/35724] [Medline: 35852842]
12. Martín A, Ashish A, Paul B, Eugene B, Zhifeng C, Craig C, et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. arXiv 2016:1-19.
13. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. NeurIPS Proceedings. URL: https://proceedings.neurips.cc/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf [accessed 2023-02-15]

14. Liu D, Sahu R, Ignatov V, Gottlieb D, Mandl KD. High Performance Computing on Flat FHIR Files Created with the New SMART/HL7 Bulk Data Access Standard. AMIA Annu Symp Proc 2019;2019:592-596 [FREE Full text] [Medline: 32308853]
15. Oehm J, Storck M, Fechner M, Brix T, Yildirim K, Dugas M. FhirExtinguisher: A FHIR Resource Flattening Tool Using FHIRPath. Stud Health Technol Inform 2021 May 27;281:1112-1113. [doi: 10.3233/SHTI210369] [Medline: 34042862]
16. Zhou L, Suominen H, Gedeon T. Adapting State-of-the-Art Deep Language Models to Clinical Information Extraction Systems: Potentials, Challenges, and Solutions. JMIR Med Inform 2019 Apr 25;7(2):e11499 [FREE Full text] [doi: 10.2196/11499] [Medline: 31021325]
17. Zong N, Wen A, Stone DJ, Sharma DK, Wang C, Yu Y, et al. Developing an FHIR-Based Computational Pipeline for Automatic Population of Case Report Forms for Colorectal Cancer Clinical Trials Using Electronic Health Records. JCO Clinical Cancer Informatics 2020 Nov(4):201-209. [doi: 10.1200/cci.19.00116]
18. Mittelstadt BD, Floridi L. The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts. Sci Eng Ethics 2016 Apr;22(2):303-341. [doi: 10.1007/s11948-015-9652-2] [Medline: 26002496]
19. Denney MJ, Long DM, Armistead MG, Anderson JL, Conway BN. Validating the extract, transform, load process used to populate a large clinical research database. Int J Med Inform 2016 Oct;94:271-274 [FREE Full text] [doi: 10.1016/j.ijmedinf.2016.07.009] [Medline: 27506144]
20. Johnson A, Bulgarelli L, Pollard T, Horng S, Celi LA, Mark R. MIMIC-IV. PhysioNet 2021;101(23):e215-e220. [doi: 10.13026/s6n6-xd98]
21. Islam N. FHIR® Resources. GitHub. URL: https://github.com/nazrulworld/fhir.resources [accessed 2023-02-15]
22. PostgreSQL: The World's Most Advanced Open Source Relational Database. PostgreSQL Global Development Group. URL: https://www.postgresql.org [accessed 2023-02-15]
23. McKinney W. Data Structures for Statistical Computing in Python. SciPy.org. 2010. URL: https://conference.scipy.org/proceedings/scipy2010/pdfs/mckinney.pdf [accessed 2023-02-14]
24. Hong N, Wen A, Shen F, Sohn S, Wang C, Liu H, et al. Developing a scalable FHIR-based clinical data normalization pipeline for standardizing and integrating unstructured and structured electronic health record data. JAMIA Open 2019 Dec;2(4):570-579 [FREE Full text] [doi: 10.1093/jamiaopen/ooz056] [Medline: 32025655]
25. Konečný J, McMahan H, Ramage D, Richtárik P. Federated Optimization: Distributed Machine Learning for On-Device Intelligence. arXiv 2016 Oct 8:1-38 [FREE Full text]
26. Ulrich H, Behrend P, Wiedekopf J, Drenkhahn C, Kock-Schoppenhauer A, Ingenerf J. Hands on the Medical Informatics Initiative Core Data Set - Lessons Learned from Converting the MIMIC-IV. Stud Health Technol Inform 2021 Sep 21;283:119-126. [doi: 10.3233/SHTI210549] [Medline: 34545827]
27. Bennett A, Wiedekopf J, Ulrich H, Johnson A. MIMIC-IV Clinical Database Demo on FHIR (version 2). PhysioNet 2022:e-215-e-220.
28. Shimabukuro DW, Barton CW, Feldman MD, Mataraso SJ, Das R. Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomised clinical trial. BMJ Open Respir Res 2017;4(1):e000234 [FREE Full text] [doi: 10.1136/bmjresp-2017-000234] [Medline: 29435343]
29. Sarmiento R, Dernoncourt F. Improving Patient Cohort Identification Using Natural Language Processing. Data MITC 2016:405. [doi: 10.1007/978-3-319-43742-2_28] [Medline: 31314253]
30. Li J, Carayon P. Health Care 4.0: A Vision for Smart and Connected Health Care. IISE Trans Healthc Syst Eng 2021 Feb 15;11(3):171-180 [FREE Full text] [doi: 10.1080/24725579.2021.1884627] [Medline: 34497970]

Abbreviations

AI: artificial intelligence
DB: database
DHP: data harmonization pipeline
EHR: Electronic Health Record
FHIR: Fast Healthcare Interoperability Resources
FL: federated learning
MIMIC: Medical Information Mart for Intensive Care

Edited by C Lovis; submitted 07.11.22; peer-reviewed by A Bartschke, K Gupta; comments to author 27.12.22; revised version received 24.01.23; accepted 25.01.23; published 21.03.23.

Please cite as:
Williams E, Kienast M, Medawar E, Reinelt J, Merola A, Klopfenstein SAI, Flint AR, Heeren P, Poncette AS, Balzer F, Beimes J, von Bünau P, Chromik J, Arnrich B, Scherf N, Niehaus S
A Standardized Clinical Data Harmonization Pipeline for Scalable AI Application Deployment (FHIR-DHP): Validation and Usability Study
JMIR Med Inform 2023;11:e43847
URL: https://medinform.jmir.org/2023/1/e43847
doi:10.2196/43847
PMID:36943344

Original Paper

A SNOMED CT Mapping Guideline for the Local Terms Used to Document Clinical Findings and Procedures in Electronic Medical Records in South Korea: Methodological Study

Sumi Sung[1], DPhil; Hyeoun-Ae Park[2], DPhil; Hyesil Jung[3], DPhil; Hannah Kang[4], DPhil

[1]Biomedical Research Institute, Seoul National University Hospital, Seoul, Republic of Korea

[2]College of Nursing, Seoul National University, Seoul, Republic of Korea

[3]Department of Nursing, Inha University, Incheon, Republic of Korea

[4]Kakao Healthcare Corp, Seongnam-si, Gyeonggi-do, Republic of Korea

Corresponding Author:
Hyeoun-Ae Park, DPhil
College of Nursing
Seoul National University
103 Daehak-ro
Jongno-gu
Seoul, 03080
Republic of Korea
Phone: 82 2 740 8827
Fax: 82 2 765 4103
Email: hapark@snu.ac.kr

Abstract

Background: South Korea joined SNOMED International as the 39th member country. To ensure semantic interoperability, South Korea introduced SNOMED CT (Systemized Nomenclature of Medicine–Clinical Terms) in 2020. However, there is no methodology to map local Korean terms to SNOMED CT. Instead, this is performed sporadically and independently at each local medical institution. The quality of the mapping, therefore, cannot be guaranteed.

Objective: This study aimed to develop and introduce a guideline to map local Korean terms to the SNOMED CT used to document clinical findings and procedures in electronic health records at health care institutions in South Korea.

Methods: The guidelines were developed from December 2020 to December 2022. An extensive literature review was conducted. The overall structures and contents of the guidelines with diverse use cases were developed by referencing the existing SNOMED CT mapping guidelines, previous studies related to SNOMED CT mapping, and the experiences of the committee members. The developed guidelines were validated by a guideline review panel.

Results: The SNOMED CT mapping guidelines developed in this study recommended the following 9 steps: define the purpose and scope of the map, extract terms, preprocess source terms, preprocess source terms using clinical context, select a search term, use search strategies to find SNOMED CT concepts using a browser, classify mapping correlations, validate the map, and build the final map format.

Conclusions: The guidelines developed in this study can support the standardized mapping of local Korean terms into SNOMED CT. Mapping specialists can use this guideline to improve the mapping quality performed at individual local medical institutions.

(JMIR Med Inform 2023;11:e46127) doi:10.2196/46127

KEYWORDS

semantic interoperability; Systematized Nomenclature of Medicine–Clinical Terms; mapping guideline; local terms; mapping; guideline; SNOMED; nomenclature; interoperable; interoperability; terminology; medical term; health term; terminologies; ontologies

Introduction

South Korea is a leader in global information and communication technology. According to a 2021 Organization for Economic Co-operation and Development (OECD) survey of national health data infrastructure and governance, South Korea ranked second among OECD countries on data availability, maturity, and use [1]. However, data are not used fully, owing to a lack of interoperability and data security problems [2]. Interoperability is the ability of different information systems, devices, and applications to access, exchange, integrate, and cooperatively use data in a coordinated manner [3]. This occurs within and across organizational, regional, and national boundaries. Interoperability provides timely and seamless use of information and helps to globally improve the health of individuals and populations [3]. There are 4 levels of interoperability: foundational, structural, semantic, and organizational. Among them, the key strategy for ensuring semantic interoperability is the use of standard terminology that allows concepts to be represented unambiguously between the senders and receivers of information [4].

To achieve semantic interoperability, interface or local terms extracted from natural language written by a health care provider can be stored as reference terminology, such as in the SNOMED CT (Systemized Nomenclature of Medicine–Clinical Terms; SNOMED International). The stored terms with reference terminology can then be used in classification systems such as the International Classification of Diseases (ICD) for statistical purposes [2]. Health care providers in South Korea write medical records in natural language rather than using standard interface terms. Korean Standard Classification of Disease (KCD) codes are used for mortality and morbidity reports, and electronic data interchange (EDI) codes are used for national health insurance claims. Therefore, to fully use health care data, mapping terms extracted from phrases written in natural languages or using interface terms in medical records, disease classification codes, and national health insurance claim codes to reference terminologies are required.

Various efforts have mapped the terms used to document clinical findings and procedures in electronic medical records (EMRs), classification systems such as ICD 10th revision, and existing health care terminologies such as Logical Observation Identifiers Names and Codes (LOINC) and International Classification of Nursing Practice (ICNP) to the SNOMED CT in other countries [5-11]. The South Korean government, aiming to encourage the use of standard terminology in health care institutions, joined SNOMED International as the 39th member country and introduced SNOMED CT in 2020 to ensure interoperability. Subsequently, various efforts have mapped terms used to document clinical findings and procedures in EMRs or national health checkup questionnaires, classification systems including KCD-7, and EDI codes to the SNOMED CT [12-17]. Furthermore, these results are used for research purposes in the Common Data Model [17,18].

Individual medical institutions have attempted to map their terms to SNOMED CT. However, the mapping quality cannot be guaranteed due to its sporadic and independent map development. SNOMED International introduced Snap2, a tool to support mapping. However, since the tool is based on English source terms, it is difficult to apply to the Korean terms used in South Korea. This study, therefore, aimed to develop a guideline to ensure high-quality mapping of terms used to document clinical findings and procedures in the EMRs of local institutions in South Korea to SNOMED CT. The guideline focuses on a process of defining a relationship between concepts used in EMRs and the concepts of SNOMED CT [18]. The guideline does not include organizing a mapping team or reviewing existing maps. In addition, this guideline's scope is limited to mapping to the SNOMED CT concept and excludes mapping to SNOMED CT post-coordinated expression.

Methods

The process of developing the mapping guideline was led by a mapping guideline development committee. The committee consisted of 3 mapping experts. All committee members had SNOMED CT mapping experiences spanning more than 5 years and have conducted various national projects, such as SNOMED CT mapping of KCD-7 and EDI codes supported by the Korea Health Information Service under the Ministry of Health and Welfare in South Korea. The development of the guidelines was conducted from December 2020 to December 2022.

To develop the mapping guidelines, the committee members first developed an overall structure based on a review of the existing guidelines and their own mapping and teaching experiences. An extensive literature review was performed in PubMed, MEDLINE, and Google Scholar. The existing mapping guidelines and previous studies [19-24] were reviewed according to these criteria: (1) scope and purpose, (2) involvement of the stakeholders, (3) rigor of development, (4) clarity of presentation, and (5) applicability [25]. After the review, the existing mapping guidelines and previous studies could not be used due to the following reasons: (1) not matched in scope or purpose because of mapping SNOMED CT concepts to other classification systems such as ICD-10 or ICD-9-CM [20,21], (2) limited to automatic mapping only without any information about mapping rules [21,22], (3) lack of currency in SNOMED CT, published more than 10 years ago [20,21], and (4) no detailed information on the mapping process [23,24]. As a result, version 2.0 of the SNOMED CT-AU mapping guideline, developed by the National Electronic Health Transition Authority of Australia, also known as the Australia Digital Health Agency [19], was chosen as the framework of the guideline. The SNOMED CT-AU mapping guideline clearly describes the preprocessing process, classification and validation of mapping results, and final map structures. It matches the scope and purpose of our SNOMED CT mapping guideline and uses the most recent version of SNOMED CT as a target code among the existing guidelines. In addition, it includes rigorous mapping examples. The following sections were adopted from SNOMED CT-AU: define the purpose and scope of the map, preprocessing source terms, mapping patterns, validation, and structure of the map.

Based on the committee members' experiences, the most difficult aspects of mapping between local terms and SNOMED CT were extracting and understanding the source terms, developing search terms, and searching the target SNOMED CT. However, the SNOMED CT-AU mapping guideline does not describe how to extract and understand source terms from EMRs, how to develop search terms, or how to search for target concepts using the SNOMED International browser. It is difficult for beginners to apply the existing SNOMED CT mapping guidelines, such as those of the SNOMED CT-AU. Therefore, the committee added 4 steps to the guideline: extract terms, preprocess source terms using clinical context, select a search term, and use search strategies to find SNOMED CT concepts using a browser.

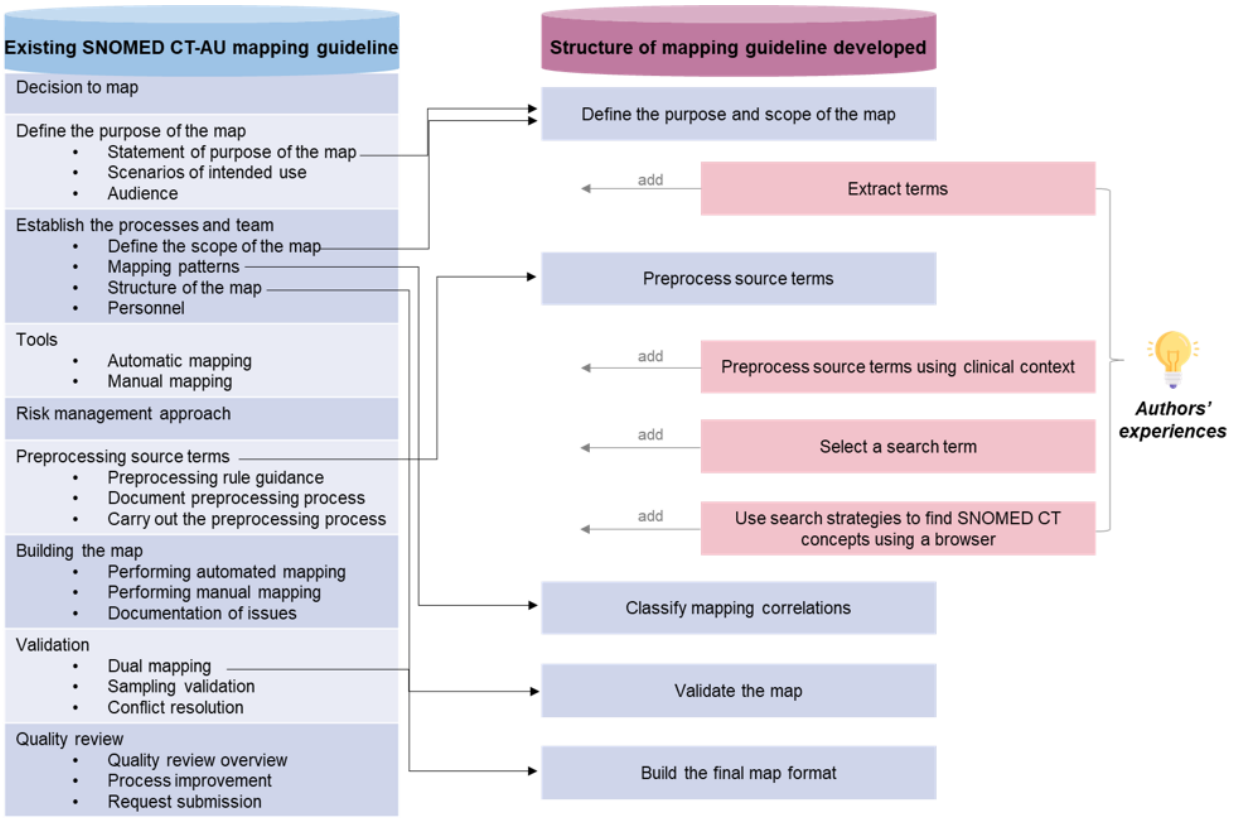
A guideline review panel was invited to validate the guideline from May 2021 to September 2022. The review panel consisted of 3 mapping experts: a professor with more than 5 years of SNOMED CT mapping experience working at a South Korean university and 2 mapping specialists with more than 3 years of mapping experience. The first author emailed, explaining the purpose of the study, and asked to review the understandability and usability of the developed guidelines and to add more mapping examples, if possible, to the panel. In addition, 2 graduate students with no mapping experience were asked to assess whether the guideline was understandable and helpful throughout the email. These 2 processes were repeated until no more issues were identified.

Results

Overview

The SNOMED CT mapping guidelines developed in this study consisted of 9 steps, as presented in Figure 1. Step 1 defines the purpose and scope of the map. Step 2 describes how to extract source terms from EMRs. Step 3 explains how to translate the source terms into Korean, how to define abbreviations or acronyms, and how to correct spelling or punctuation errors [19]. Step 4 explains how to understand the meanings of the extracted source terms to reflect clinical contexts. Step 5 describes how to select appropriate search terms to improve mapping [26]. Step 6 describes efficient strategies to search for SNOMED CT concepts using a browser [27]. Step 7 describes how to classify the correlation between source terms and target SNOMED CT [5,7,28-31]. Step 8 explains how to validate the adequacy and accuracy of the map [19]. The final step explains how to document the final map [32]. The SNOMED CT examples used in this guideline are taken from the international edition, released on January 31, 2023.

Figure 1. Overview of the developed guideline structure. SNOMED CT: Systemized Nomenclature of Medicine–Clinical Terms.



Step 1: Define the Purpose and Scope of the Map

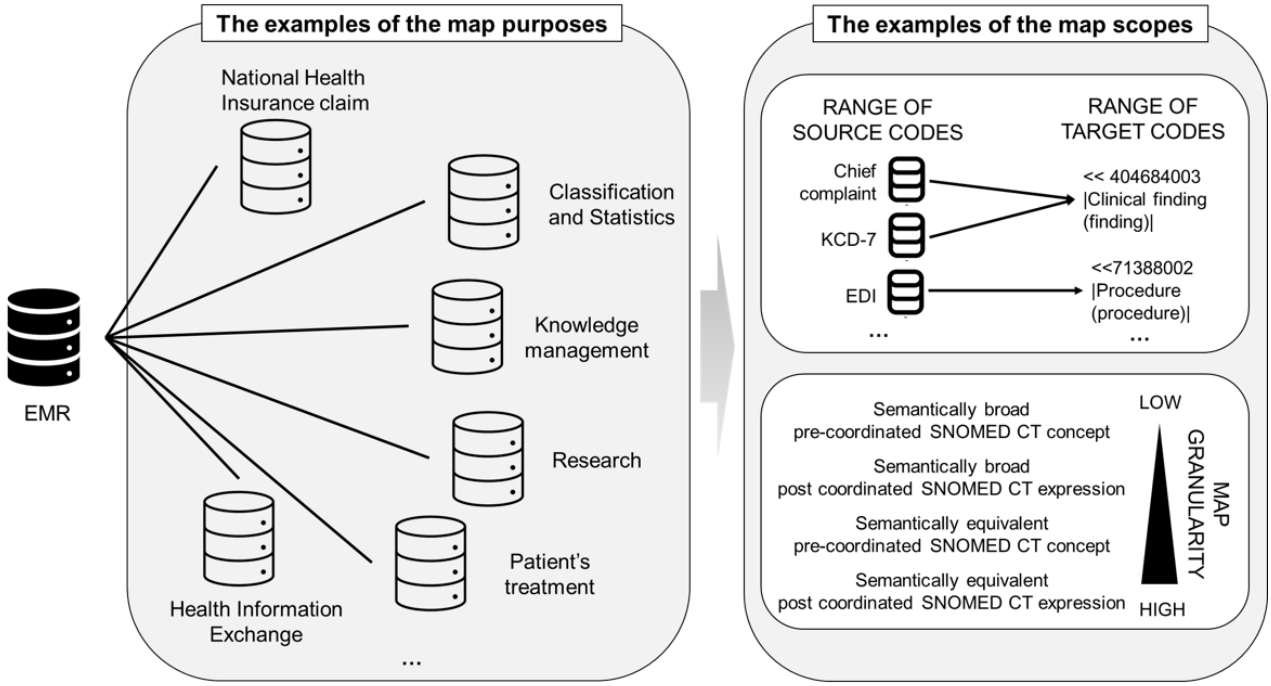
To proceed with mapping, the mapping specialists must first define the purpose and scope. The purpose of the map can be for national health insurance claims, classification and statistics (eg, mortality and prevalence of specific diseases), knowledge management (eg, decision support system), health information exchange among institutions, studies (eg, Common Data Model), and patient care. Figure 2 presents examples of map purposes and scopes.

The scope of the map refers to the range of source and target codes and granularity and may differ according to its purpose. For example, if the map's purpose is to provide prevalence

statistics for specific diseases, the source codes are restricted to KCD-7 in the clinical findings domain, and the range of target codes should be SNOMED CT concepts in the clinical findings, person, or event top-level hierarchy. If the map's purpose is for national health insurance claims, the source codes are restricted to the EDI codes in the procedure domain, and the range of target codes should be SNOMED CT concepts in the procedures' top-level hierarchy.

Map granularity can also vary according to the purpose of the map. For example, if the map is for national health insurance claims, the local term "alcohol-related seizure," mapped to KCD-7 code G40.5 (special epileptic syndromes), should be mapped to the abstract SNOMED CT concept 230431001 (situation-related seizures [disorder]). However, if the map is for exchanging health information among institutions, the source code should be mapped at the granular level to SNOMED CT concept 308742005 (alcohol withdrawal-induced convulsion [disorder]).

Figure 2. Examples of map purpose and scope. EDI: electronic data interchange; EMR: electronic medical record; KCD: Korean Classification of Disease; SNOMED CT: Systemized Nomenclature of Medicine–Clinical Terms.



Step 2: Extract Terms

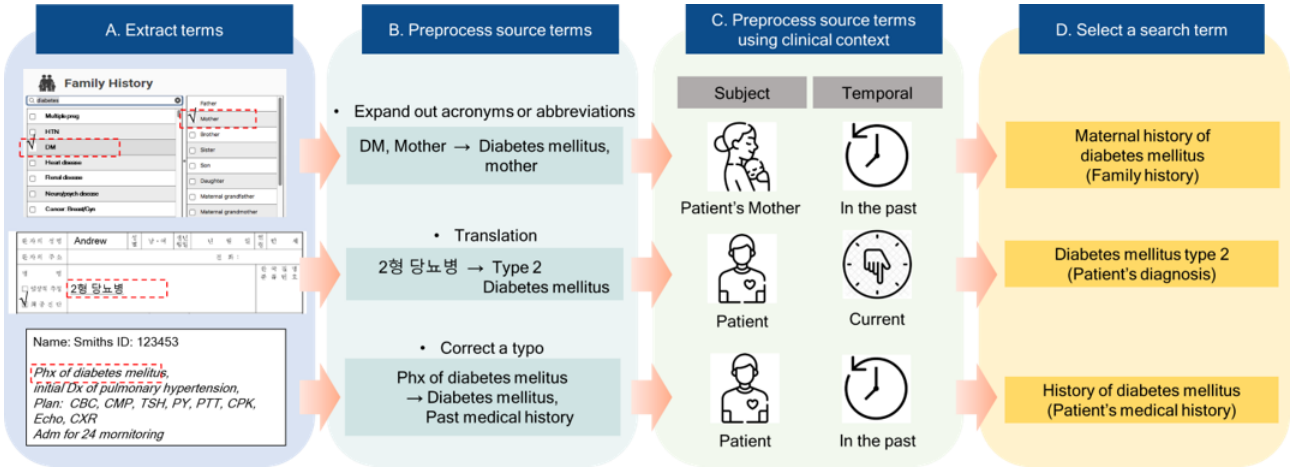
When the purpose and scope of the map are determined, mapping specialists should extract terms from EMRs while considering where and how clinical notes are documented. Patient diagnoses can be extracted from progress notes and other similar information sources. Family and past medical histories can be extracted from progress notes and initial nursing assessment records. Surgical procedure names can be extracted from surgical notes. Evaluation procedures can be extracted from laboratory documentation.

These data sources are either structured, semistructured, or unstructured medical records written in natural language; for example, as presented in step A of Figure 3, if we extract a "patient's mother diabetes mellitus type 2," first the interface terms from the family medical history sections in EMRs can be identified, and then the "DM" diagnosis and the "mother's family relationship" can be extracted. The terms written in free text from the semistructured family medical history records sections of EMRs can be identified, and then the diagnosis "2형 당뇨병 (diabetes mellitus, type 2)" can be extracted. We also can identify terms from unstructured records by reading all the documents and extracting the terms written in free text, such as "Past medical history of diabetes mellitus," which requires natural language processing (NLP).

When conducting automatic term extraction through NLP, word segmentation throughout semantic analysis is required using a corpus [33,34]. For example, the term "diabetes mellitus" should be extracted based on its meaning, not by extracting "diabetes" and "mellitus" separately. For this process to be possible, a Korean medical corpus can be used, such as the medical terminology database released by the Korean Medical Association [35] or the Korean Medical Library Engine [36]. The results of automatic terminology extraction should be reviewed manually.

Alternatively, terms can simply be extracted from code systems such as KCD-7, EDI, and local hospital codes. In these cases, it is sufficient to extract the terms mapped to the code system without duplication.

Figure 3. Examples of the process from term extraction to selecting a search term.



Step 3: Preprocess the Source Terms

Preprocessing takes daily English terms as the source terms, for instance, by translating Korean terms into English. These Korean English (Konglish) terms are then translated into native English terms, with acronyms or abbreviations defined and spelling or punctuation errors corrected. If terms are written in Korean, preprocessing starts with translating the Korean terms into English. For example, "당뇨" written in Korean must be translated into "diabetes mellitus." Konglish is often used in Korean clinical settings. It is necessary to change Konglish terms into proper English terms. In Korean clinical settings, for example, an evaluation procedure "초음파 검사" is represented by various English words such as in SONO and sonogram. SONO, or sonogram, is the image obtained using ultrasound. If it is the name of an evaluation procedure, it is, therefore, appropriate to translate it into "ultrasound."

If terms are written in acronyms or abbreviations, which are commonly used by health care providers, preprocessing may be required to rewrite them in complete words. The acronym ASD, used in departments of pediatric cardiology, was defined as "atrial septal defect." The abbreviation MMG used frequently in South Korea, must be rewritten to "mammography."

Terms should be corrected when misspelled. For example, "asceding aorta dilatation" should be corrected to "ascending aorta dilatation" to obtain the correct search results. Furthermore, if terms include incorrect punctuation, it is preferable to edit the spacing of the source terms. For example, the search for the source term "DeQuervain's disease, wrist, Rt." is missing a space, which must be added to produce "De Quervain." Other examples are presented in step B of Figure 3.

Step 4: Preprocess Source Terms Using Clinical Context

The terms preprocessed in the previous step can have different meanings depending on the clinical context, such as subject, temporal, or finding contexts. For example, the meaning of the preprocessed source term "diabetes mellitus" varies depending on which part of the EMR it is extracted from; it can be "diabetes mellitus," "past medical history of diabetes mellitus," "family history of diabetes mellitus," or "no family history of diabetes mellitus." "Past medical history of diabetes mellitus" means that the temporal context is in the past. "Family history of diabetes mellitus" means that the temporal context is current or in the past and that the subject is a person in the patient's family. "No family history of diabetes mellitus" means that the temporal and subject contexts are the same as "family history of diabetes mellitus," but the finding context is absent.

Clinical context can be inferred from the source of the structured or semistructured medical records or the text in which the source term is written. If the family history section is the source of the medical record, the subject is a family member of the patient, and the temporal context is the past or the present. If the past medical history section is the source, the subject is the patient, and the temporal context is the past or present. If the chief complaint section is the source, the subject is the patient, and the temporal context is the present. If the clinical context cannot be inferred from the structured or semistructured records, the mapping personnel should read the text in which the source term is written. Examples are presented in step C of Figure 3.

Step 5: Select a Search Term

To search for a target SNOMED CT concept, a search term comprising keywords of the source terms that have undergone preprocessing must be selected. The search success rate is increased if the mapping specialist understands the general naming convention used by SNOMED International [26]. For example, a diagnosis or chief complaint can be expressed as "finding or morphology + body structure," a surgery or procedure name can be expressed as "procedure + body structure," the past medical history can be expressed as "history + disease name," and the family medical history can be expressed as "family history + disease name." The body structure requires "structure" to be attached to its name. Most descriptions used singular rather than plural expressions.

Stop words in NLP, which are frequently used words with restricted semantic specificity, should be excluded from search terms to improve the success rate. These stop words include articles ("a," "an," "the"), prepositions ("in," "at," "with," "without"), conjunctions ("and," "as"), and ambiguous adjectives or adverbs ("other," "alone," "single," "side," etc). Examples are presented in step D of Figure 3.

Step 6: Use Search Strategies to Find SNOMED CT Concepts Using a Browser

Use the First 3 Characters of Words

Searches can start with the first 3 characters of search terms. With the first 3 characters, the mapping specialist can prevent search failures due to differences in the forms of various words, such as nouns, verbs, and the use of the passive voices and differences in singular and passive voices. For example, if the source term "perforated small intestine" is entered into the browser, equivalently matched concepts cannot be obtained. However, if 3 characters of each word—"per sma int"—are entered, the concept "235741002 |Perforation of small intestine (disorder)|" is matched.

When a SNOMED CT concept appears, the parent, child, and sibling concepts—higher, lower, or at the same level in the hierarchy—must be reviewed to determine if it is semantically correct.

Search With Synonyms of Source Terms

If the target SNOMED CT concept is not identified by search terms, synonyms should be used as search terms for the source term. The synonyms can be obtained by searching the medical terminology database released by the Korean Medical Association [34] or by searching the Korean Medical Library Engine [36]. Otherwise, the synonym can be found by including "synonym" in search terms on search engines such as Google, searching the thesaurus, reviewing websites of prestigious medical institutions, or from previous studies. In addition, conversational English terms should be changed so that different terms with equivalent meanings are also considered. For example, the conversational English phrase "removal of prostate stones" should be converted into the medically accurate phrase "removal of prostate calculus" for a successful search. Table 1 lists examples of using synonyms in a search term. If a concept is searched using synonyms, the parent, child, and sibling concepts must be reviewed in the final process to confirm whether the concept is semantically correct.

Table 1. Examples of using synonyms.

Source term	Search term
Removal of prostate *stones*	Removal of prostate *calculus*
Abdomen *sonogram*	*Ultrasonography* of abdomen
Operation of nystagmus	Nystagmus *surgery*
Extraction of nail	*Removal* of nail
Correction for rectocele	*Repair* of rectocele
Cervical mucus *test*	Cervical mucus *analysis*

Start With Broad or Narrow Terms

If the SNOMED CT concept is not found with synonyms of search terms, the target concept that is semantically equivalent to the source term can be searched by starting with broader or narrower terms. For example, "IUD ectopic" in the diagnosis means that an intrauterine device is in an abnormal place or position. If "intrauterine device ectopic" is entered as a search term, no search results are obtained. If "intrauterine device" (which is a broad term for a concept in the diagnosis) is searched for and the results are filtered through the "disorder" semantic tag, 16 disorder-related terms are retrieved. Among them, the "malposition of intrauterine contraceptive device" concept is semantically equivalent to the "IUD ectopic" concept. In this process, the parent, child, and sibling concepts must be reviewed to confirm whether the concept is semantically correct, as mentioned above in the final process.

Filter SNOMED CT Concept Using Semantic Tag

Critically, to map semantically equivalent target SNOMED CT concepts, the mapping specialist must filter results by semantic tag in the SNOMED CT browser when entering a search term. These tags are presented in Table 2. For example, if the source term "cold" in the diagnosis is entered in the browser without filters, 2 lexically matched concepts 82272006 (common cold [disorder]) and 84162001 (cold sensation quality [qualifier value]) with the same description "cold" are obtained. However, if the results are filtered through the semantic tag "disorder," the semantically equivalent concept, 82272006 (common cold [disorder]), is immediately obtained. The diagnosis "renal cell carcinoma" must be filtered through the "disorder" semantic tag, and mapped to 702391001 (renal cell carcinoma [disorder]), not 41607009 (renal cell carcinoma [morphologic abnormality]).

Table 2. SNOMED CT semantic tags according to the clinical domain of the source codes.

Clinical domain of source concepts	SNOMED CT[a] semantic tag
Chief complaint	finding
Diagnosis	disorder, finding
Past medical history, Family medical history	situation
Health-related behavior history	finding
Complication	finding
Procedure: (surgical procedure; evaluation procedure)	procedure, regime/therapy, observable entity
Pathologic diagnosis	morphologic abnormality
Medication	clinical drug, medicinal product, medicinal product form, substance

[a]SNOMED CT: Systematized Nomenclature of Medicine–Clinical Terms.

Step 7: Classify Mapping Correlations

The mapping results can be classified according to the correlation between the source concepts and SNOMED CT concepts or according to map cardinality, depending on the number of concepts.

Mapping Classification According to the Correlation

According to the correlation, the mapping results can be classified as "1193548004 |Exact match between map source and map target (foundation metadata concept)|," "1193549007 |Narrow map source to broad map target (foundation metadata concept)|," or "1193551006 |Map source not mappable to map target (foundation metadata concept)|." When the source term is matched to the equivalent SNOMED CT concept, the map is classified as a "1193548004 |Exact match between map source and map target (foundation metadata concept)|." When the source term is matched to broader SNOMED CT concepts, the map is classified as a "1193549007 |Narrow map source to broad map target (foundation metadata concept)|." When no concept broadly matches a source term, the map is classified as "1193551006 |Map source not mappable to map target (foundation metadata concept)|."

As an example, the source term "breast cancer, upper inner quadrant" in a diagnosis is equivalently matched to 373082000 (malignant neoplasm of breast upper inner quadrant [disorder]), and the map is classified as "1193548004 |Exact match between map source and map target (foundation metadata concept)|." Another example is that the source term "aortic valve stenosis occurred after mitral valve replacement" in the diagnosis has no equivalent SNOMED CT concept that can be equivalently mapped, so it can be mapped to the broadly matching "703223000 |Postprocedural aortic valve stenosis (disorder)|." In this case, the map is classified as "1193549007 |Narrow map source to broad map target (foundation metadata concept)|."

Classify Mapping According to Map Cardinality

If a single source code does not map to a single SNOMED CT concept, it is necessary to classify the map's cardinality using a complex map. The mapping results can be classified as either "one to one" or "one to many" according to their cardinality. When a single source code is mapped to a single SNOMED CT concept, the map is classified as a "one to one." When a single source code with broad or multiple meanings is mapped to multiple SNOMED CT concepts, it is classified as "one to many." For example, the source term "right or left hemicolectomy" has 2 meanings: "right hemicolectomy" and "left hemicolectomy." The source term is mapped to "359571009 |Right colectomy (procedure)|" and "82619000 |Left colectomy (procedure)|," respectively, and the map is classified as "one to many."

Step 8: Validate the Map

The map is validated with internal and external validation. Internal validation may vary depending on how many individuals participated in the mapping and validation processes. When 2 mapping experts are involved, one maps the source terms and the other reviews the mapping results. Otherwise, both mapping experts could map the same source terms and then compare the maps constructed by each other. When more than 2 mapping experts are involved, a validation can be conducted by dividing them into 2 groups—mapping and reviewing groups. The map is deemed to be correct if the mappers and reviewers select the same results. If the maps differ, the results should be evaluated in a group discussion with the other mapper to agree on which SNOMED CT concept to use.

External validation can also be used, in which clinical or mapping experts who were not involved in the mapping process verify the validity of the mapping results. Methods for performing external validation include reviewing the sample mapping results and obtaining mapping results that were difficult to map or that were not agreed upon in an internal group discussion.

Step 9: Build the Final Map Format

There are 2 types of maps—simple and complex maps. A simple map is a representation of mapping from a term in other code systems to a SNOMED CT concept; this is comprised of the source code (id, term), target SNOMED CT (id, fully specified name), and map correlation. A complex map is a representation of mapping from a term in other code systems to one or more SNOMED CT concepts; this is comprised of the source code (id, term), target SNOMED CT (id, fully specified name), map correlation, and cardinality. If map cardinality is 2 or more, rows are added to the final map. When documenting the final map, versions of source codes and SNOMED CTs, mapping

dates, and mapper information must be included. Figure 4 presents examples of the mapping format.

Figure 4. Examples of final map formats. SNOMED CT: Systemized Nomenclature of Medicine–Clinical Terms.

[Simple map]

Source code		Target SNOMED CT		Map correlation	Mapping date	Mapper
ID	Term	ID	Fully Specified Name			
B02	Zoster [herpes zoster]	4740000	Herpes zoster (disorder)	1193548004 |Exact match between map source and map target (foundation metadata concept)|	YYYYMMDD	SS, HJ
B05	Measles	14189004	Measles (disorder)	1193548004 |Exact match between map source and map target (foundation metadata concept)|	YYYYMMDD	SS, HJ
A18	Tuberculosis of other organs	56717001	Tuberculosis	1193549007 |Narrow map source to broad map target (foundation metadata concept)|	YYYYMMDD	HJ, HK

[Complex map]

Source code		Target SNOMED CT		Map correlation	Map Cardinality	Mapping date	Mapper
ID	Term	ID	Fully Specified Name				
D00	Carcinoma in situ of oral cavity, oesophagus and stomach	78579800 2	Carcinoma in situ of oral cavity (disorder)	1193548004 |Exact match between map source and map target (foundation metadata concept)|	3	YYYYMMDD	SS, HJ
		92585006	Carcinoma in situ of esophagus (disorder)				
		92756002	Carcinoma in situ of stomach (disorder)				

Discussion

Principal Results

A SNOMED CT mapping guideline has been developed for the terms used to document clinical findings and procedures in EMRs. It was developed based on a review of previous mapping guidelines and the literature and the experiences of the authors and the map guideline development committee members. During the development of this mapping guideline, we reflected on the methods of resolving the difficulties experienced in KCD-7 and EDI code mapping, such as preprocessing the source terms and selecting search terms. We also reflected on the methods and examples identified while teaching mapping specialists working in local institutions to map the terms used to document clinical findings and procedures in EMRs to the SNOMED CT at the Korea Human Resource Development Institute of South Korea since 2019 [37]. The mapping training program was evaluated as having high satisfaction among mapping specialists with an average of 4.25 out of 5 points in 2021 (1=very low to 5=very high) and was reported as helpful in conducting mapping during work when they returned to their job with 3.75 out of 5 points in 2021 (1=not helpful to 5=very helpful). The guideline developed in this study will therefore be useful for local mapping specialists to map the terms used to document clinical findings and procedures in EMRs to the SNOMED CT. The guideline will improve the mapping quality performed at each medical institution.

The guidelines developed in this study contain detailed mapping steps from the Korean perspective, not only in the new steps such as term extraction, syntactic and semantic preprocessing, search term selection, and searching strategies in the SNOMED International browser but also in the steps adopted from previous mapping guidelines and studies [19,27,38]. For example, South Korean EMRs are written in both English and Korean. When written in Korean, translation is mandatory during preprocessing, and even when written in English, the Korean way of expressing a medical concept in English differs from that in English-speaking countries, so additional preprocessing is also required. Since Korean mapping specialists are limited by their knowledge of English synonyms, examples of synonyms were included to make the guideline easier to understand. The guideline can, therefore, be applied in local medical institutions to map Korean terms used to document clinical findings and procedures in EMRs to SNOMED CT. They can also be used to develop SNOMED CT mapping guidelines in other countries.

Limitations

This study had some limitations. Since the guideline focused on specific clinical domains, other clinical domains, such as medicine, were not included. Future studies are required on the development of mapping guidelines for terms used to document other clinical domains. In addition, this guideline does not include a guideline for mapping to SNOMED CT post-coordinated expression. The post-coordinated expressions frequently used in South Korea can be added as a new concept in the Korean extension of SNOMED CT.

Conclusions

This study developed a SNOMED CT mapping guideline for the terms used to document clinical findings and procedures in EMRs at local institutions in South Korea. The guideline was based on existing mapping guidelines, the findings of previous studies, and the mapping and teaching experiences of the authors. The mapping guideline developed in this study consisted of the following nine steps: (1) define the purpose and scope of the map, (2) extract terms, (3) preprocess source terms, (4) preprocess source terms using clinical context, (5) select a search term, (6) use search strategies to find SNOMED CT concepts using a browser, (7) classify mapping correlations, (8) validate the map, and (9) build the final map format. The new guideline can be published on the website of the Korea Health Information Service. The guideline can be applied to local medical institutions when mapping Korean terms used to document clinical findings and procedures in EMRs to SNOMED CT. It will also support local medical institutions in standardizing their local code systems using SNOMED CT. Ultimately, the data quality of each local medical institution will be improved, allowing the data to be fully used in clinical

decision support systems, health information exchange, and clinical research.

Acknowledgments

This study was supported by the Korea Health Information Service, and by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare and Ministry of Science and ICT, Republic of Korea (grant HU22C0168).

Conflicts of Interest

None declared.

References

1. Oderkirk J. Survey results: national health data infrastructure and governance. OECD Health Working Papers. 2021. URL: https://tinyurl.com/mr3vcank [accessed 2023-04-06]
2. Park HA, Yu SJ, Jung H. Strategies for adopting and implementing SNOMED CT in Korea. Healthc Inform Res 2021;27(1):3-10 [FREE Full text] [doi: 10.4258/hir.2021.27.1.3] [Medline: 33611871]
3. Interoperability in healthcare. Healthcare Information and Management Systems Society. URL: https://www.himss.org/resources/interoperability-healthcare [accessed 2021-10-26]
4. Park HA, Hardiker N. Clinical terminologies: a solution for semantic interoperability. J Kor Soc Med Inform 2009;15(1):1-11. [doi: 10.4258/jksmi.2009.15.1.1]
5. Thandi M, Brown S, Wong ST. Mapping frailty concepts to SNOMED CT. Int J Med Inform 2021;149:104409. [doi: 10.1016/j.ijmedinf.2021.104409] [Medline: 33677397]
6. Lougheed MD, Thomas NJ, Wasilewski NV, Morra AH, Minard JP. Use of SNOMED CT and LOINC to standardize terminology for primary care asthma electronic health records. J Asthma 2018;55(6):629-639. [doi: 10.1080/02770903.2017.1362424] [Medline: 28800265]
7. Block L, Handfield S. Mapping wound assessment data elements in SNOMED CT. Stud Health Technol Inform 2016;225:1078-1079. [Medline: 27332492]
8. Bodenreider O. Issues in mapping LOINC laboratory tests to SNOMED CT. AMIA Annu Symp Proc 2008;2008:51-55 [FREE Full text] [Medline: 18999311]
9. Elkin PL, Brown SH, Husser C, Wahner-Roedler D, Bailey S, Nugent L, et al. Using SNOMED-CT as a reference terminology: mapping VA disability terminology to ICD-9-CM. Connecting Medical Informatics and Bio-Informatics. 2005. URL: https://tinyurl.com/yc2fdrk9 [accessed 2023-04-05]
10. Kim TY, Hardiker N, Coenen A. Inter-terminology mapping of nursing problems. J Biomed Inform 2014;49:213-220 [FREE Full text] [doi: 10.1016/j.jbi.2014.03.001] [Medline: 24632297]
11. Nguyen AN, Truran D, Kemp M, Koopman B, Conlan D, O'Dwyer J, et al. Computer-assisted diagnostic coding: effectiveness of an NLP-based approach using SNOMED CT to ICD-10 mappings. AMIA Annu Symp Proc 2018;2018:807-816 [FREE Full text] [Medline: 30815123]
12. KCD-SNOMED CT Mapping Table 2022. Korea Health Information Service. URL: https://www.hins.or.kr/termMapping/kcdMappingList.es?mid=a11301020000 [accessed 2022-04-06]
13. EDI-SNOMED CT mapping table: Korea health information service. Korea Health Information Service. URL: https://www.hins.or.kr/termMapping/ediMappingList.es?mid=a11301030000 [accessed 2022-04-06]
14. So EY, Park HA. Mapping medical records of gastrectomy patients to SNOMED CT. Stud Health Technol Inform 2011;169:764-768. [Medline: 21893850]
15. So EY, Park HA. Exploring the possibility of information sharing between the medical and nursing domains by mapping medical records to SNOMED CT and ICNP. Healthc Inform Res 2011;17(3):156-161 [FREE Full text] [doi: 10.4258/hir.2011.17.3.156] [Medline: 22084810]
16. Hwang JE, Park HA, Shin SY. Mapping the Korean national health checkup questionnaire to standard terminologies. Healthc Inform Res 2021;27(4):287-297 [FREE Full text] [doi: 10.4258/hir.2021.27.4.287] [Medline: 34788909]
17. Kang H, Park HA. Mapping Korean national health insurance reimbursement claim codes for therapeutic and surgical procedures to SNOMED-CT to facilitate data reuse. Stud Health Technol Inform 2022;290:101-105. [doi: 10.3233/SHTI220040] [Medline: 35672979]
18. Health Informatics — Principles of Mapping Between Terminological Systems. ISO/TR. 2014. URL: https://www.iso.org/standard/51344.html [accessed 2023-04-05]
19. SNOMED CT-AU: mapping guidelines v2.0. National Electronic Health Transition Authority. 2014. URL: https://developer.digitalhealth.gov.au/specifications/ehealth-foundations/ep-2372-2016/nehta-1790-2014 [accessed 2023-04-05]
20. Imel M, Campbell JR. Mapping from a clinical terminology to a classification. AHIMA's 75th Anniversary National Convention and Exhibit Proceedings. 2003. URL: https://library.ahima.org/doc?oid=61537 [accessed 2023-04-05]

21. Giannangelo K, Millar J. Mapping SNOMED CT to ICD-10. Stud Health Technol Inform 2012;180:83-87. [doi: 10.1007/978-1-4471-2801-4_15]
22. Fung KW, Xu J, Ameye F, Gutiérrez AR, D'Havé A. Leveraging lexical matching and ontological alignment to map SNOMED CT surgical procedures to ICD-10-PCS. AMIA Annu Symp Proc 2016;2016:570-579 [FREE Full text] [Medline: 28269853]
23. Osornio AL, Luna D, Gambarte ML, Gomez A, Reynoso G, de Quirós FG. Creation of a local interface terminology to SNOMED CT. Stud Health Technol Inform 2007;129(Pt 1):765-769. [Medline: 17911820]
24. Bakhshi-Raiez F, Ahmadian L, Cornet R, de Jonge E, de Keizer NF. Construction of an interface terminology on SNOMED CT. Generic approach and its application in intensive care. Methods Inf Med 2018;49(4):349-359. [doi: 10.3414/me09-01-0057]
25. Brouwers MC, Kho ME, Browman GP, Burgers JS, Cluzeau F, Feder G, AGREE Next Steps Consortium. AGREE II: advancing guideline development, reporting, and evaluation in health care. Prev Med 2010;51(5):421-424. [doi: 10.1016/j.ypmed.2010.08.005] [Medline: 20728466]
26. SNOMED CT Editorial Guide - General 2022. International Health Terminology Standards Organisation. URL: https://confluence.ihtsdotools.org/display/DOCEG/SNOMED+CT+Editorial+Guide+-+General [accessed 2022-07-21]
27. SNOMED CT search and data entry guide. SNOMED CT International. 2017. URL: https://confluence.ihtsdotools.org/display/DOCSEARCH [accessed 2023-04-08]
28. Hwang EJ, Park HA, Sohn SK, Lee HB, Choi HK, Ha S, et al. Mapping Korean EDI medical procedure code to SNOMED CT. Stud Health Technol Inform 2019;264:178-182. [doi: 10.3233/SHTI190207] [Medline: 31437909]
29. Lee DH, Lau FY, Quan H. A method for encoding clinical datasets with SNOMED CT. BMC Med Inform Decis Mak 2010;10(1):53 [FREE Full text] [doi: 10.1186/1472-6947-10-53] [Medline: 20849611]
30. Peters L, Kapusnik-Uner JE, Bodenreider O. Methods for managing variation in clinical drug names. AMIA Annu Symp Proc 2010;2010:637-641 [FREE Full text] [Medline: 21347056]
31. Zhou L, Plasek JM, Mahoney LM, Chang FY, DiMaggio D, Rocha RA. Mapping partners master drug dictionary to RxNorm using an NLP-based approach. J Biomed Inform 2012;45(4):626-633 [FREE Full text] [doi: 10.1016/j.jbi.2011.11.006] [Medline: 22142948]
32. SNOMED CT Release File Specifications. 2022. URL: http://snomed.org/rfs [accessed 2023-04-04]
33. Velupillai S, Mowery D, South BR, Kvist M, Dalianis H. Recent advances in clinical natural language processing in support of semantic analysis. Yearb Med Inform 2015;10(1):183-193 [FREE Full text] [doi: 10.15265/IY-2015-009] [Medline: 26293867]
34. Hongying Z, Wenxin L, Kunli Z, Yajuan Y, Baobao C, Zhifang S. Building a pediatric medical corpus: word segmentation and named entity annotation. In: Chinese Lexical Semantics Workshop. Cham: Springer; 2020:21-664.
35. Medical terminology database. Korean Medical Association. URL: https://term.kma.org/search/list.asp [accessed 2023-03-02]
36. Korean Medical Library Engine. URL: http://www.kmle.co.kr/ [accessed 2023-03-02]
37. SNOMED CT education and training in South Korea. SNOMED CT Expo 2022. URL: https://confluence.ihtsdotools.org/display/FT/SNOMED+CT+Expo+2022 [accessed 2023-01-18]
38. Højen AR, Rosenbeck Gøeg KR. SNOMED CT implementation. Mapping guidelines facilitating reuse of data. Methods Inf Med 2018;51(06):529-538. [doi: 10.3414/me11-02-0023]

Abbreviations

EDI: electronic data interchange
EMR: electronic medical record
ICD: International Classification of Diseases
ICNP: International Classification of Nursing Practice
KCD: Korean Standard Classification of Disease
LOINC: Logical Observation Identifiers Names and Codes
NLP: natural language processing
OECD: Organization for Economic Co-operation and Development
SNOMED CT: Systemized Nomenclature of Medicine–Clinical Terms

Edited by Q Chen; submitted 31.01.23; peer-reviewed by C Gaudet-Blavignac, T Kang, X Jiang, D Lee; comments to author 15.02.23; revised version received 06.03.23; accepted 30.03.23; published 18.04.23.

Please cite as:
Sung S, Park HA, Jung H, Kang H
A SNOMED CT Mapping Guideline for the Local Terms Used to Document Clinical Findings and Procedures in Electronic Medical Records in South Korea: Methodological Study
JMIR Med Inform 2023;11:e46127
URL: https://medinform.jmir.org/2023/1/e46127
doi:10.2196/46127
PMID:37071456

Viewpoint

Practical Considerations for Developing Clinical Natural Language Processing Systems for Population Health Management and Measurement

Suzanne Tamang[1,2], PhD; Marie Humbert-Droz[1], PhD; Milena Gianfrancesco[3], PhD; Zara Izadi[3], PhD; Gabriela Schmajuk[3], MD; Jinoos Yazdany[3], PhD

[1]Division of Immunology and Rheumatology, Stanford University School of Medicine, Stanford, CA, United States
[2]Department of Veterans Affairs, Office of Mental Health and Suicide Prevention, Program Evaluation Resource Center, Palo Alto, CA, United States
[3]Division of Rheumatology, University of California, San Francisco, San Francisco, CA, United States

Corresponding Author:
Jinoos Yazdany, PhD
Division of Rheumatology
University of California, San Francisco
10 Koret Way, Room K-219
San Francisco, CA, 94143
United States
Phone: 1 415 576 1000
Email: jinoos.yazdany@ucsf.edu

Abstract

Experts have noted a concerning gap between clinical natural language processing (NLP) research and real-world applications, such as clinical decision support. To help address this gap, in this viewpoint, we enumerate a set of practical considerations for developing an NLP system to support real-world clinical needs and improve health outcomes. They include determining (1) the readiness of the data and compute resources for NLP, (2) the organizational incentives to use and maintain the NLP systems, and (3) the feasibility of implementation and continued monitoring. These considerations are intended to benefit the design of future clinical NLP projects and can be applied across a variety of settings, including large health systems or smaller clinical practices that have adopted electronic medical records in the United States and globally.

(JMIR Med Inform 2023;11:e37805) doi:10.2196/37805

KEYWORDS

clinical natural language processing; electronic health records; population health science; clinical decision support; information extraction

Introduction

Natural Language Processing (NLP) has the potential to improve the delivery, quality, and safety of health care [1-7]. There have been numerous research applications, including the extraction of disorders, drugs, and procedures. Moreover, NLP methods have automated the extraction of information that is likely to be undercoded or not coded in a patient's record, such as the severity of a disorder, their functional status, or social determinants of health [4,6-8]. However, examples of health systems operationalizing clinical NLP tools for real-world clinical decision-making, as well as population health management and quality measurement, are limited. This is a missed opportunity to turn rich, unstructured data into structured information that can be used for quality and performance initiatives within a health system or a professional field, or to make national-level comparisons [2,9-12].

To address the challenges translating research tools to clinical practice, we present practical considerations for NLP system stakeholders that can be used to position an early-stage research project for use in real-world decision-making and to eventually demonstrate institutional value. Our practical considerations are informed by prior literature and reports that describe a chiasm rather than a synergy between clinical NLP research and clinical practice. For example, Wen et al [13] share the Mayo Clinic's Desiderata for the implementation of an NLP development delivery platform derived from 2 decades of implementing clinical NLP in their health system. Lederman et al [14] describe how existing clinical NLP systems "have delivered marginal practical utility and are rarely deployed into

health care settings" and call for a new paradigm of clinical NLP research for real-world decision support. Similarly, Newman-Griffis et al [15] call for a new paradigm and general principles for clinical NLP research that are focused on challenges posed by application needs and describe how these challenges can drive innovation in basic science and technology design. Referring to artificial intelligence systems in medicine more broadly, Topol et al [16] have also observed that "deployment of medical AI systems in routine clinical care provides an important yet largely unfulfilled opportunity". We also draw from our own collective experience developing clinical NLP systems for research studies and in an operational capacity.

Our practical considerations can be used to support the development of applications that can push forward advances in clinical medicine right now. We also assess the current landscape of Clinical NLP tools and techniques on our adjoining public GitHub site, which can be updated by the research community as clinical NLP technologies evolve [17].

Practical Consideration #1: Are Data and Compute Infrastructure Ready for NLP?

"Garbage in, garbage out" refers to low-quality data, or "garbage," that can result in misinformation. It was first used by US Army scientists to provide the intuition that computers cannot think for themselves, and that "sloppily programmed" inputs inevitably lead to incorrect outputs. Although this saying is over a half century old, it applies even more today, when powerful computers can record large amounts of data that are not fit for the intended use in a short amount of time.

Key questions that will help to determine NLP readiness of a new clinical corpus includes the following: (1) Are notes and note metadata reported in a timely way and with reasonable quality? (2) Is the unstructured free-text data ready for NLP techniques (eg, can the data be used to extract clinical concepts with an accuracy that is fit for the indented use)? And (3) Are the NLP algorithms feasible to execute in the production environment?

Assessing the quality of textual data—or "Data Readiness"—confronts on the problem of data quality by providing empirical findings about syntactic and semantic aspects of a clinical corpus as well as the associated note metadata such as patient identifiers, the date and time of the note, and the type of note. We define "quality" within the context defined by Kahn et al [18] where three dimensions are considered, including plausibility, conformance, and completeness. The results of a Data Readiness assessment help to predict the difficulty of building an NLP system for those data. The quality of free-text data can vary significantly between different note types within the same or across different electronic medical record (EMR) systems. For example, discharge summaries typically contain complete sentences and clearly demarcated sections. By contrast, intensive care unit (ICU) progress notes typically contain large quantities of digits that are not explicitly labelled as to whether they are vital signs, ventilator settings, or any of the many other quantitative measures that are monitored in critically ill patients. ICU progress notes also frequently contain large amounts of information in just one or two grammatically unstructured sentences. Ambulatory progress reports can range from a few sentences to longer documents with standardized formats.

In some cases, data sets that do not initially appear to be ready for NLP on an intended task can be further processed or sampled so that the data are more amenable to their intended use. For example, a data source can be preprocessed to remove notes that do not fit predetermined plausibility criteria, such as the known range of system availability to identify notes that have a plausible date, or notes that do not have an indicated date. However, this may not always result in data that are ready for NLP; in these cases, investigators should work with an organization leadership to improve data collection before undertaking an NLP project.

The institutional nuances of EMR clinical documentation processes require clinical NLP systems developed at other institutions to be customized to a new local data set. This uses specific preprocessing steps related to the provenance and structure of the source data. In prior work funded by the Agency for Healthcare Research and Quality that was based on the Rheumatology Informatics System for Effectiveness (RISE) registry, we found that simple summary statistics on note length in characters and words ("tokens") were helpful to assess the quality of clinical notes from rheumatology practices across the United States [12]. The RISE registry began operation in 2014, and the free-text extraction covered the period between 2014 and 2018. It combines data from over 260 ambulatory outpatient rheumatology practices that collectively use more than 20 different EMR products. To assess the data readiness of RISE for health services research and to better understand the epidemiology of chronic rheumatic diseases, we first used note metadata. For example, we calculated the number of unique clinical notes recorded by year, as indicated by the time stamp of the patient note. Unique notes were determined by each entry of a textual document within the RISE database. We found that many notes had an invalid timestamp, with dates as early as 1800 and as far out as 8018. This suggested an opportunity to improve the quality of these data. We also found that simple summary statistics describing the textual data helped determine the potential informativeness of RISE for scientific and practical applications. Table 1 suggests that RISE contains many relatively short patient notes (mean of 34.57 tokens in 2018) as well as some longer, more traditional patient notes and letters (SD 203.01 tokens). These types of summary statistics are an important first step in NLP data readiness assessments.

Table 1. Mean, SD, minimum, mode, and maximum note length[a] and word count[b] for free-text patient Rheumatology Notes submitted to the American College of Rheumatology's data registry, by year.

Year	Note count	Length of note				Word count			
		Mean (SD)	Min[c]	Mode	Max[d]	Mean (SD)	Min	Mode	Max
2010	891,837	96 (353)	1	17	18,774	16 (54)	__[e]	2	2549
2011	1,238,711	128 (554)	4	17	40,295	20 (80)	1	2	5713
2012	2,412,737	118 (559)	3	19	23,370	18 (81)	1	2	3496
2013	3,409,806	120 (597)	3	19	23,921	18 (87)	1	2	3567
2014	5,394,083	209 (1069)	1	19	614,356	31 (160)	—	2	107,498
2015	7,715,894	211 (1547)	1	19	2,179,227	31 (224)	—	2	375,620
2016	9,812,735	233 (1356)	1	19	425,503	34 (186)	—	2	75,844
2017	11,685,000	242 (1468)	1	19	570,721	35 (204)	—	2	100,311
2018	5,301,039	239 (1415)	1	19	192,570	35 (203)	—	2	31,852
Total	50,222,840	205 (1271)	1	19	2,179,227	30 (180)	—	2	375,620

[a]Please note that 2018 is a partial year. Note length is indicated by non–whitespace characters and symbols.

[b]Word count was estimated after deidentification of the Rheumatology Informatics System for Effectiveness corpus.

[c]Min: minimum.

[d]Max: maximum.

[e]Not available.

To assess the readiness of data for specific linguistic analysis tasks, such as part-of-speech tagging or named-entity recognition, there are a variety of other descriptive statistics based on corpus linguistics that can be used to assess the quality of textual data. Some of these focus on gross characteristics of the data, such as the extent to which documents have clearly identifiable sections and the nature of the data in those sections. For example, lists such as diagnoses and medications usually have relatively clear boundaries, while family and individual medical histories may not. The presence or absence of sentence boundaries, as well as the length of sentences, are also important predictors of the effort required to build high-performing language processing tools. Other descriptive statistics assess textual data on the level of individual words. For example, textual genres with high levels of repeated word use (eg, fever and pain) can be easier to process than textual genres with high levels of words that only appear once (eg, misspellings and typographic errors).

In addition to data that are ready for NLP, automated information extraction algorithms require infrastructure that will allow for the efficient processing of large volumes of new patient notes. There must be discussions at the design phase of the project to ensure that any research products can be operationally tested, and if warranted, translated to operational infrastructure. It is also important for the product to be updated and maintained if being used longitudinally with routine updates of notes.

If a project has no feasible pathway to operationalize the NLP system for real-world decision support, it might be possible that new resources, including institutional computing infrastructure, could be recommended and acquired.

Practical Consideration #2: What are the Incentives for Adopting the NLP System?

Key questions that will help determine if the proper incentives are in place to support a Clinical NLP system are as follows: (1) will the NLP help to address an existing clinical need? (2) is there support from clinical leadership for the ongoing use of the NLP system? and (3) is there a financial incentive to adopting the NLP system?

Reporting from structured data has been the mainstay in health care practice for decades. The Sentinel active surveillance system for medical products and Observational Medical Outcomes Partnership (OMOP) initiatives helped to pioneer the use of common data models to support regulatory initiatives [19,20]. Building on OMOP's common data models, the Observation Health Data Science Initiative's extension has extended the OMOP schema to incorporate unstructured data with the "NOTE" and "NOTE_NLP" tables. It is likely that EMR databases will become even more powerful for regulatory initiatives when they can jointly leverage various data modalities such as patient notes or images for the purpose of improved patient care. However, in the absence of a specific clinical need that a system is designed to address, and without the proper incentives to use the system, it is unlikely that a system will be adopted for clinical uses such as decision support, regardless of performance on a research task. A successful system for population and precision health must be innovative, pragmatic enough to be deployed in a production environment and directly aligned with organizational incentives and clinical leadership's priorities. It should support interoperability but also allow for customization to the nuances of different health systems. We discuss some of these challenges in the next section.

In cases where there is little or no organizational incentive to adopt a clinical NLP system, it is unlikely to succeed past the research phase. Therefore, working with leadership to identify the potential value to a health system and finding possible incentives to adopt such a system are important first steps.

Practical Consideration #3: Feasibility of Implementation and Evaluation

Key questions that will help to determine the feasibility of implementing and evaluating a clinical NLP system include the following: (1) What is the task (ie, clinical need) that this system seeks to address? (2) Are the clinical concepts of interest captured in structured data? If so, are there limitations to what can be extracted? (3) If NLP is justified, are simple NLP techniques enough or are more complex algorithms warranted? (4) Can the Clinical NLP tool be developed and implemented in a reasonable timeline to fulfill stakeholder needs? (5) What are potential sources of bias, considering factors such as the NLP approach, the data used to train the NLP tool, and the population to which it is applied?

An important early consideration is regarding the target population. In cross-validation over random folds, models are trained and tested over the same population. However, in practice, models are often developed in a training data set but applied to novel data that may originate from a different underlying population of patients or clinicians. Differences in clinical practice and workflow patterns, as well as lack of homogeneity in clinical language (as described above), can have large impacts on the transportability of models from where they were developed to a given target population. This is important to factor into the training assessment (eg, being aware of overfitting) and possibly also into model development. If an available external test set exists that represents the target population, it should be tested as part of the model development process to ensure that the NLP tool is portable and externally valid. Ideally, performance metric reporting should be required for all tools meant to be transportable outside of their training corpus.

There are multiple strategies for mitigating bias and improving portability of NLP tools. One source of bias may arise from the specific type of note used to develop a model; for example, an NLP tool developed only on ICU notes, pathology reports, or notes within a certain specialty may not generalize to other note types or clinical settings. Therefore, different note types should be incorporated into the training corpus, if in fact, the target corpus is intended to involve multiple types. Additionally, as previously described, incorporating a secondary data set that represents the target population for testing, apart from the primary data set used for training, can help ensure that the model is transportable and performs well across health care settings, EMRs, and patient populations.

To evaluate model performance, one must decide at which level the assessment should occur, that is, at the mention, document, or patient level. NLP models can be evaluated by their precision (positive predictive value), recall (sensitivity), specificity, F_1-score (harmonic mean of precision and recall) and overall accuracy compared to a "gold-standard" test set of reviewed text [5,9]. However, the text-specific evaluation may not be as important as the document or even patient-level performance, especially if multiple mentions per patient occur, or structured data fields are being incorporated into the evaluation in conjunction with NLP annotations. Therefore, although at the mention level, the NLP model may correctly identify a patient as positive, it may be that it is only when combined with the additional information (other mentions, lab results, etc) that the output and model performance are clinically important.

As important as model performance at the time of development is, more crucial may be the model performance over time. Validation of NLP models is key both retrospectively and prospectively, as data change longitudinally. It is important for models to be evaluated continually to determine whether they should be fine-tuned and updated, and whether any biases exist. For example, this may involve updating rule-based code to reflect changes in language representation or reevaluating or redeveloping deep learning–based NLP models.

If a clinical NLP system does not address a known and ideally high-priority clinical need, it is less likely to be adopted into practice. However, it may be possible to adapt the system to address a need identified by organizational leadership. If it does not initially show good performance, continued development may help to improve the clinical systems accuracy, especially if linguistic annotation data can be generated and made available for training a better model. Lastly, in some cases where additional expertise can be used for a project, it may be possible to meet a project deadline that would otherwise not be possible. Importantly, having a strategy, including a business plan, for maintaining deployed models is important to ensuring that their clinical application is sustained.

The Potential Role of NLP in Real-World Decision Support

NLP has the potential to improve population health outcomes in the United States. For example, in the inpatient care setting, NLP systems could reliably identify individuals with symptoms of diarrhea reported in progress notes and feed these data into algorithms for *Clostridioides difficile* testing. Inpatients with falls documented in clinical notes could trigger alerts to discontinue sedatives or narcotics. In the outpatient setting, NLP can be used to assess the severity of a disease or a postoperative complication. The NLP of free-text patient notes also creates opportunities for national, routine quality and performance measurement, which can support improvement in the value of health care delivered to patients at highest risk for poor outcomes [9-12,21-23]. As health systems across the United States move toward whole-person care paradigms, NLP systems can also be used to identify important clinical decision support factors that are undercoded or altogether absent from structured data sources in patient records, such as the presence of behavioral, psychosocial, and economic risk factors.

Predictive analytics is another area where incorporating clinical text has the potential to improve population health [5-7,24]. Most models for population-level risk stratification that use

health care data have exclusively relied on structured data, but several groups have demonstrated that in certain domains, adding information from clinical text can improve performance. Studies in this area reflect a wide range of tasks from predicting hospital readmissions to identifying patients at risk for suicide [2-13,17,21,22,24,25]. Such models can be used operationally to more accurately target a subset of a population for specific interventions designed to address modifiable risk factors.

Applications of NLP to streamline and facilitate quality and safety reporting are also emerging [9-12]. Federal reporting of quality and safety measures often places considerable burden on clinicians, sometimes requiring duplicate entry of similar concepts in the text of clinical notes as well as in structured fields that can be queried to calculate performance. Reliable extraction of relevant information from clinical notes would not only alleviate burdensome data entry but also greatly expand the types of concepts included in reporting programs. For example, guidelines in rheumatology support the routine collection of disease activity scores for patients with rheumatoid arthritis, but not all EMRs have structured fields to input these scores. Electronic quality measures that extract this information automatically from structured fields might miss scores that are documented only in clinical notes. NLP could be used to extract these scores and improve the validity and reliability of such quality measures[9-12].

While these and other applications of NLP have the potential to improve health care and population health, the successful deployment and dissemination of these applications has been limited. Given these barriers, how should the field move forward? In addition to our three considerations, we think it is critical that multiple stakeholders provide input from the start of NLP projects. Practicing clinicians can ensure the focus of the work is clinically relevant, fulfills an unmet need, and is aligned with current clinical workflows; clinical informaticians can provide insight into whether data systems are available to scale valid NLP algorithms, and health care administrators can lend insight into IT resources required and the feasibility of scaling and sustaining systems. Until there is stakeholder alignment and investment in a project, impact and scalability are likely to be limited. Similar to many new technologies in medicine, alignment often requires the development of the NLP program as a value proposition that either clearly impacts operational efficiency, revenue, quality and safety, or patient outcomes. Moreover, stakeholders need to be integrated into the software development life cycle to ensure the product's ongoing implementation is successful.

Conclusion

The analysis of unstructured free-text patient data enables new ways in which scientific questions can be studied and health care can be delivered. Although such uses are promising, leveraging the clinical text data collected in the EMR and using these data in health care operations are not without substantial caveats. Opportunities to better align state-of-the-art systems developed by researchers to support the measurement of patient-reported outcomes and to support high-quality health care delivery can likely lead to improved outcomes. With a focus on designing practical applications that are aligned with clinical requirements and organizational incentives, the considerations listed here can be used to design a project-specific checklist for a variety of stakeholders. We also summarized the procedures for considering appropriate use of NLP in health and survey the current landscape of Clinical NLP tools. To support future work in this area, we have provided software and data set summaries, license, and other access requirements on our adjoining GitHub site, which we hope will serve as a continuously updated resource for the research community as technologies evolve.

Conflicts of Interest

None declared.

References

1. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. Nat Med 2019 Jan 7;25(1):24-29. [doi: 10.1038/s41591-018-0316-z] [Medline: 30617335]
2. Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated Encoding of Clinical Documents Based on Natural Language Processing. J Am Med Inform Assoc 2004 Sep 01;11(5):392-402. [doi: 10.1197/jamia.m1552]
3. Masanz J, Pakhomov S, Xu H, Wu S, Chute C, Liu H. Open Source Clinical NLP - More than Any Single System. AMIA Jt Summits Transl Sci Proc 2014;2014:76-82 [FREE Full text] [Medline: 25954581]
4. Wang J, Deng H, Liu B, Hu A, Liang J, Fan L, et al. Systematic Evaluation of Research Progress on Natural Language Processing in Medicine Over the Past 20 Years: Bibliometric Study on PubMed. J Med Internet Res 2020 Jan 23;22(1):e16816 [FREE Full text] [doi: 10.2196/16816] [Medline: 32012074]
5. Liu F, Weng C, Yu H. Advancing Clinical Research Through Natural Language Processing on Electronic Health Records: Traditional Machine Learning Meets Deep Learning. In: Clinical Research Informatics. Cham, Switzerland: Springer International Publishing; 2019:357-378.
6. Wu S, Roberts K, Datta S, Du J, Ji Z, Si Y, et al. Deep learning in clinical natural language processing: a methodical review. J Am Med Inform Assoc 2020 Mar 01;27(3):457-470 [FREE Full text] [doi: 10.1093/jamia/ocz200] [Medline: 31794016]
7. Yang Z, Dehmer M, Yli-Harja O, Emmert-Streib F. Combining deep learning with token selection for patient phenotyping from electronic health records. Sci Rep 2020 Jan 29;10(1):1432 [FREE Full text] [doi: 10.1038/s41598-020-58178-1] [Medline: 31996705]

8. Hao T, Huang Z, Liang L, Weng H, Tang B. Health Natural Language Processing: Methodology Development and Applications. JMIR Med Inform 2021 Oct 21;9(10):e23898 [FREE Full text] [doi: 10.2196/23898] [Medline: 34673533]
9. Murff HJ, FitzHenry F, Matheny ME, Gentry N, Kotter KL, Crimin K, et al. Automated identification of postoperative complications within an electronic medical record using natural language processing. JAMA 2011 Aug 24;306(8):848-855. [doi: 10.1001/jama.2011.1204] [Medline: 21862746]
10. Yetisgen M, Klassen P, Tarczy-Hornoch P. Automating data abstraction in a quality improvement platform for surgical and interventional procedures. EGEMS (Wash DC) 2014 Nov 26;2(1):1114 [FREE Full text] [doi: 10.13063/2327-9214.1114] [Medline: 25848598]
11. Tamang S, Hernandez-Boussard T, Ross E, Gaskin G, Patel M, Shah N. Enhanced Quality Measurement Event Detection: An Application to Physician Reporting. EGEMS (Wash DC) 2017 May 30;5(1):5 [FREE Full text] [doi: 10.13063/2327-9214.1270] [Medline: 29881731]
12. Humbert-Droz M, Izadi Z, Schmajuk G, Gianfrancesco M, Baker MC, Yazdany J, et al. Development of a Natural Language Processing System for Extracting Rheumatoid Arthritis Outcomes From Clinical Notes Using the National Rheumatology Informatics System for Effectiveness Registry. Arthritis Care Res (Hoboken). Preprint posted online March 14, 2022. [doi: 10.1002/acr.24869] [Medline: 35157365]
13. Wen A, Fu S, Moon S, El Wazir M, Rosenbaum A, Kaggal VC, et al. Desiderata for delivering NLP to accelerate healthcare AI advancement and a Mayo Clinic NLP-as-a-service implementation. NPJ Digit Med 2019;2:130 [FREE Full text] [doi: 10.1038/s41746-019-0208-8] [Medline: 31872069]
14. Lederman A, Lederman R, Verspoor K. Tasks as needs: reframing the paradigm of clinical natural language processing research for real-world decision support. J Am Med Inform Assoc 2022 Sep 12;29(10):1810-1817 [FREE Full text] [doi: 10.1093/jamia/ocac121] [Medline: 35848784]
15. Newman-Griffis D, Lehman JF, Rosé C, Hochheiser H. Translational NLP: A New Paradigm and General Principles for Natural Language Processing Research. Proc Conf 2021 Jun;2021:4125-4138 [FREE Full text] [Medline: 34179899]
16. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nat Med 2022 Jan;28(1):31-38 [FREE Full text] [doi: 10.1038/s41591-021-01614-0] [Medline: 35058619]
17. Tamang S. Practical Considerations for Clinical Natural Language Processing. GitHub. URL: https://github.com/suzytamang/practicalConsiderationsCNLP/wiki/Practical-Considerations-for-Healthcare-Natural-Language-Processing-Systems [accessed 2022-11-29]
18. Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, et al. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. EGEMS (Wash DC) 2016;4(1):1244 [FREE Full text] [doi: 10.13063/2327-9214.1244] [Medline: 27713905]
19. Observational Medical Outcomes Partnership (OMOP). FNIH. URL: https://fnih.org/what-we-do/major-completed-programs/observational-medical-outcomes-partnership-omop [accessed 2022-11-29]
20. About the Food and Drug Administration (FDA) Sentinel Initiative. Sentinel Initiative. 2016. URL: https://www.sentinelinitiative.org/about [accessed 2022-11-29]
21. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting Information from Textual Documents in the Electronic Health Record: A Review of Recent Research. Yearb Med Inform 2018 Mar 07;17(01):128-144. [doi: 10.1055/s-0038-1638592]
22. FitzHenry F, Murff HJ, Matheny ME, Gentry N, Fielstein EM, Brown SH, et al. Exploring the frontier of electronic health record surveillance: the case of postoperative complications. Med Care 2013 Jun;51(6):509-516 [FREE Full text] [doi: 10.1097/MLR.0b013e31828d1210] [Medline: 23673394]
23. Capurro D, Yetisgen M, van Eaton E, Black R, Tarczy-Hornoch P. Availability of structured and unstructured clinical data for comparative effectiveness research and quality improvement: a multisite assessment. EGEMS (Wash DC) 2014 Jul 11;2(1):1079 [FREE Full text] [doi: 10.13063/2327-9214.1079] [Medline: 25848594]
24. Sheikhalishahi S, Miotto R, Dudley J, Lavelli A, Rinaldi F, Osmani V. Natural Language Processing of Clinical Notes on Chronic Diseases: Systematic Review. JMIR Med Inform 2019 Apr 27;7(2):e12239 [FREE Full text] [doi: 10.2196/12239] [Medline: 31066697]
25. Zeng J, Gensheimer MF, Rubin DL, Athey S, Shachter RD. Uncovering interpretable potential confounders in electronic medical records. Nat Commun 2022 Mar 23;13(1):1014 [FREE Full text] [doi: 10.1038/s41467-022-28546-8] [Medline: 35197467]

Abbreviations

EMR: electronic medical record
ICU: intensive care unit
NLP: natural language processing
OMOP: Observational Medical Outcomes Partnership
RISE: Rheumatology Informatics System for Effectiveness

Edited by T Hao; submitted 08.03.22; peer-reviewed by H Mehdizadeh, JD Posada Aguilar; comments to author 08.07.22; revised version received 02.09.22; accepted 09.11.22; published 03.01.23.

Please cite as:
Tamang S, Humbert-Droz M, Gianfrancesco M, Izadi Z, Schmajuk G, Yazdany J
Practical Considerations for Developing Clinical Natural Language Processing Systems for Population Health Management and Measurement
JMIR Med Inform 2023;11:e37805
URL: https://medinform.jmir.org/2023/1/e37805
doi:10.2196/37805
PMID:36595345

Original Paper

An End-to-End Natural Language Processing Application for Prediction of Medical Case Coding Complexity: Algorithm Development and Validation

He Ayu Xu[1*], PhD; Bernard Maccari[2*], MSc; Hervé Guillain[3], MD, DrPH; Julien Herzen[2], PhD; Fabio Agri[4,5*], MBA, MD; Jean Louis Raisaro[1*], PhD

[1]Biomedical Data Science Center, Lausanne University Hospital, Lausanne, Switzerland

[2]Unit8 SA, Lausanne, Switzerland

[3]Public Health Solutions Ltd, Promasens, Switzerland

[4]Department of Administration and Finance, Lausanne University Hospital, Lausanne, Switzerland

[5]Department of Visceral Surgery, Lausanne University Hospital, Lausanne, Switzerland

[*]these authors contributed equally

Corresponding Author:
He Ayu Xu, PhD
Biomedical Data Science Center
Lausanne University Hospital
CHUV, Centre hospitalier universitaire vaudois Rue du Bugnon 21
Lausanne, 1011
Switzerland
Phone: 41 0795566886
Email: he.xu@chuv.ch

Abstract

Background: Medical coding is the process that converts clinical documentation into standard medical codes. Codes are used for several key purposes in a hospital (eg, insurance reimbursement and performance analysis); therefore, their optimization is crucial. With the rapid growth of natural language processing technologies, several solutions based on artificial intelligence have been proposed to aid in medical coding by automatically suggesting relevant codes for clinical documents. However, their effectiveness is still limited to simple cases, and it is not yet clear how much value they can bring in improving coding efficiency and accuracy.

Objective: This study aimed to bring more efficiency to the coding process to improve the selection of codes by medical coders. To achieve this, we developed an innovative multimodal machine learning–based solution that, instead of predicting codes, detects the degree of coding complexity before coding is performed. The notion of coding complexity was used to better dispatch work among medical coders to eventually minimize errors and improve throughput.

Methods: To train and evaluate our approach, we collected 2060 cases rated by coders in terms of coding complexity from 1 (simplest) to 4 (most complex). We asked 2 expert coders to rate 3.01% (62/2060) of the cases as the gold standard. The agreements between experts were used as benchmarks for model evaluation. A case contains both clinical text and patient metadata from the hospital electronic health record. We extracted both text features and metadata features, then concatenated and fed them into several machine learning models. Finally, we selected 2 models. The first used cross-validated training on 1751 cases and testing on 309 cases aiming to assess the predictive power of the proposed approach and its generalizability. The second model was trained on 1998 cases and tested on the gold standard to validate the best model performance against human benchmarks.

Results: Our first model achieved a macro–F_1-score of 0.51 and an accuracy of 0.59 on classifying the 4-scale complexity. The model distinguished well between the simple (combined complexity 1-2) and complex (combined complexity 3-4) cases with a macro–F_1-score of 0.65 and an accuracy of 0.71. Our second model achieved 61% agreement with experts' ratings and a macro–F_1-score of 0.62 on the gold standard, whereas the 2 experts had a 66% (41/62) agreement ratio with a macro–F_1-score of 0.67.

Conclusions: We propose a multimodal machine learning approach that leverages information from both clinical text and patient metadata to predict the complexity of coding a case in the precoding phase. By integrating this model into the hospital coding

system, distribution of cases among coders can be done automatically with performance comparable with that of human expert coders, thus improving coding efficiency and accuracy at scale.

(JMIR Med Inform 2023;11:e38150) doi:10.2196/38150

KEYWORDS

medical coding; natural language processing; NLP; complexity prediction; prediction; decision support; machine learning; model; clinical decision support application; multimodal modeling; coding; algorithm; documentation; health record; electronic health record; EHR; development

Introduction

Background

Medical coding [1] is the translation of health care diagnoses and procedures into standard diagnosis and procedure codes using medical classifications and controlled terminologies. It is a strategic activity for funding hospitals and, therefore, its optimization is a priority in health care systems under financial pressure. In many countries worldwide, including Switzerland, hospital funding is based on the so-called *Prospective Payment System* [2,3] mechanism. In the Swiss Prospective Payment System, for example, inpatient stays are assigned to diagnosis-related groups [4] according to diagnosis and procedure codes derived from medical documentation, and each hospital stay is paid according to the diagnosis-related group to which it is assigned. Therefore, medical coding is closely linked, on the one hand, to medical documentation, and on the other hand, to hospital revenues. In addition to establishing reimbursement claims, medical codes are used for several other goals, such as setting budgets for planned hospitalizations or evaluating the quality of care by means of indicators such as complication rates after surgery.

The diagnosis and procedure codes of a specific case (ie, inpatient stay) are derived from clinical documentation such as discharge letters, surgical reports, physicians' and nurses' notes, and laboratory and radiologic results. The International Statistical Classification of Diseases and Related Health Problems, 10th Revision (ICD-10) [5], is usually used for coding diagnoses, whereas the classification system used to code procedures can vary from country to country [6].

Codes are manually entered into a hospital information system. In Switzerland, there are >200 coding rules that govern code entry and must be applied by medical coders. The latter are health care professionals who have undergone specific training for this purpose. However, despite training, medical coding remains a complex, quickly evolving, time-consuming, and error-prone task. In our tertiary academic medical center, medical coding staff have been divided into specialty teams since 2018. In a batch of cases, 50% are distributed to a "common pot," and the other 50% are distributed to the corresponding specialty teams of medical coders. The cases in the "common pot" are distributed randomly to each team. A higher percentage of cases for the specialty teams is not envisaged for 3 reasons. First, it could lead to a loss of knowledge in general coding. Second, it could cause boredom for medical coders. Third, it will not always be possible to guarantee a sufficient number of cases for certain teams. Thus, a way to increase the efficiency of the current distribution of work without going toward a counterproductive overspecialization [7] is to force cases requiring high expertise to be assigned to experienced and specialist coders. This approach is only possible by detecting the complexity of the cases in advance before they are distributed and coded.

In recent years, artificial intelligence (AI) methods have been increasingly proposed to improve the efficiency and accuracy of medical coding. Their main goal has been to support medical coders in finding the most appropriate diagnosis and procedure codes for a given medical documentation. Conventional models, deep learning models such as convolutional neural networks and long short-term memory, and transformers have been trained and tested on automatic coding tasks using publicly available data sets in English [8-13]. Recently, this work has also been expanded to non-English corpora such as the French corpus [14,15]. In addition to the academic approach, commercial software for automatic coding has also been developed and introduced to the market. For example, commercial software such as ID SUISSE [16] applies rule-based algorithms to perform automatic coding. Their principle is to use a prebuilt dictionary of ICD-10 codes and their text labels, try to find clinical text that matches the labels, and then convert the text to ICD-10 codes. More recent tools such as Collective Thinking [17] and 360 Encompass (3M) [18] have improved the rule-based algorithms with machine learning (ML) techniques. Finally, solutions such as Sumex [19] rely on statistical methods to analyze the distributions and combinations of ICD-10 codes to identify possible inconsistencies in the coding patterns.

Despite the increasing number of available solutions, the effectiveness of automatic coding is still limited. Among the best-performing ML models, although precision can reach approximately 75%, the macro–F_1-score could only achieve 10% to 12% [12,20,21]. The results indicate that even the best models can only capture a small portion of medical codes from free text. Therefore, the improvement of medical coding using AI-assisted strategies remains an open challenge (Kaur R, unpublished data, July 2021).

Objectives

The purpose of our study was not to find a way to predict ICD-10 codes from medical records. Instead, it was to improve coding quality and efficiency by predicting coding complexity before the coding process. Our primary objective was to bring more efficiency to the coding process to improve the quality of coding by medical coders, and the means to achieve this is an innovative solution using ML. The innovation is to use ML to detect complexity, which is then used to better dispatch the work among medical coders. To the best of our knowledge, this

approach has never been used before. It allows for a more efficient distribution of cases according to coders' abilities and experience. As such, we will be able to minimize potential human errors because of random assignment and uneven distributions of coding expertise within hospitals' coding divisions or units. Eventually, by knowing the coding complexity up front, simple cases can be assigned to beginners or nonspecialist coders or AI-assisted systems to maximize their utility while complex cases for which AI-assisted tools are still inefficient are assigned to coding specialists or at least to experienced medical coders.

Depending on the amount of clinical documentation to be examined and other factors such as the length of stay or the diversity of medical specialists involved in the treatment of a patient, coding a case may be a simple or a really complex task. Once a case has been coded, it is typically easy for the person who has done so to classify the case into a complexity level, which represents the complexity of the coding activity. However, predicting the complexity level of a case up front is very time-consuming for a human coder as it requires a deep analysis of the entire documentation, which eventually is equivalent to conducting the coding process directly.

To predict the complexity of a coding task in the precoding phase in an automatic way, we used advanced natural language processing (NLP) techniques to analyze clinical texts and extract features that are predictive of the complexity of cases. We proposed an end-to-end approach that integrates the NLP and ML model into the hospital clinical data warehouse and end-user coding system. Our NLP and ML model predicts case complexity with an accuracy comparable with that achieved by expert human coders. Its beta version is currently under deployment at Lausanne University Hospital. To the best of our knowledge, we are the first to propose and develop this innovative approach.

The remainder of the paper is organized as follows. The application details are presented in the *Methods* section, and the performance and analysis are presented in the *Results* section. In the *Discussion* section, we discuss the values and importance of our application as well as the use of NLP in health care.

Methods

Ethics Approval

The Cantonal Ethics Commission for research on human beings of Canton Vaud granted a full waiver for this study given the its retrospective and quality assurance nature under Req-2022-00677.

Overview

We describe a typical medical coding workflow in Figure 1. After an inpatient (patient who is hospitalized overnight) is treated in the hospital, a discharge letter is produced. Medical coders analyze the diagnosis in the discharge letter and translate the diagnosis into International Statistical Classification of Diseases and Related Health Problems, 10th Revision (ICD-10) codes. Sometimes the coders need to refer to other clinical documents (eg, intervention protocol and laboratory reports) to translate the information accurately. The diagnosis-related group codes are computed based on the ICD-10 codes and are sent to the insurance companies for billing. The insurance companies reimburse the bills to the hospital based on the received diagnosis-related group codes. If the insurance companies find mistakes in the codes, they ask for revisions from the coding service. We provide an overview of our decision support system in Figure 2 and describe its integration into the hospital information system in Figure 3.

Figure 1. The general coding procedure in hospitals. DRG: diagnosis-related group; ICD-10: International Statistical Classification of Diseases and Related Health Problems, 10th Revision.

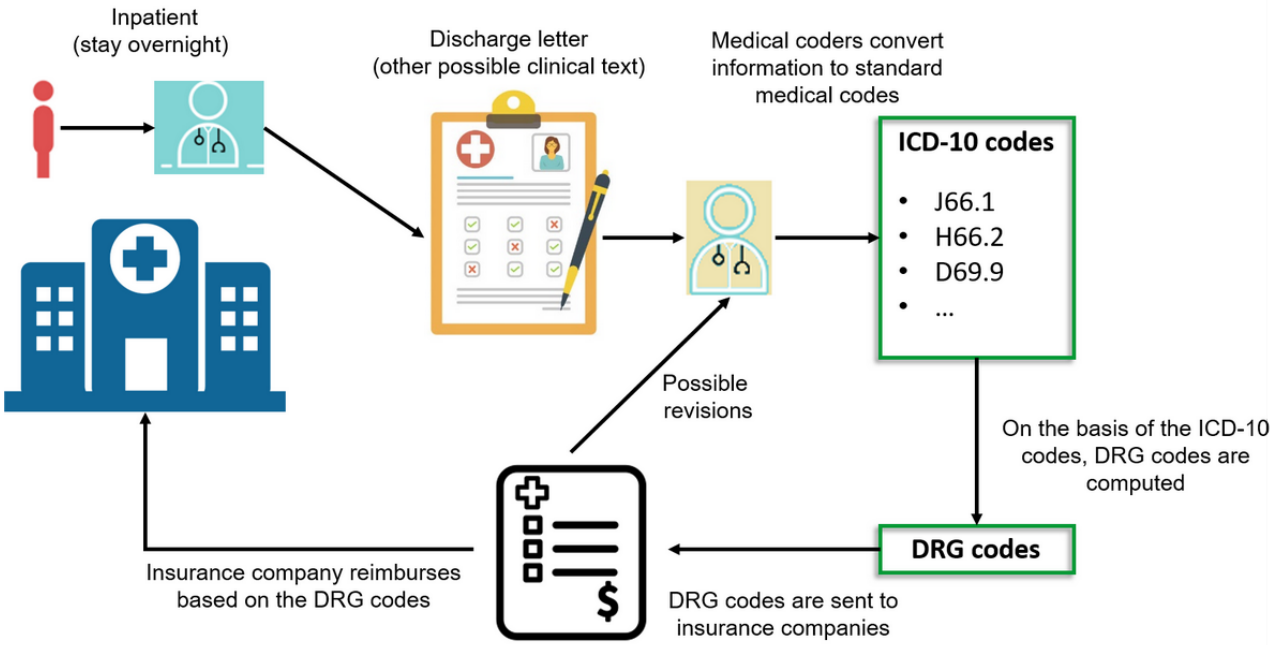


Figure 2. Workflow of this study. (A) We extracted 2060 cases from the clinical data warehouse at Lausanne University Hospital (CHUV). The cases are rated by coders (B) with complexity ranging from 1 (simplest) to 4 (most complex). (C) We performed feature engineering and trained models on the labeled cases. (D) The final model can produce both predictions of the complexity and its confidence in the predictions.

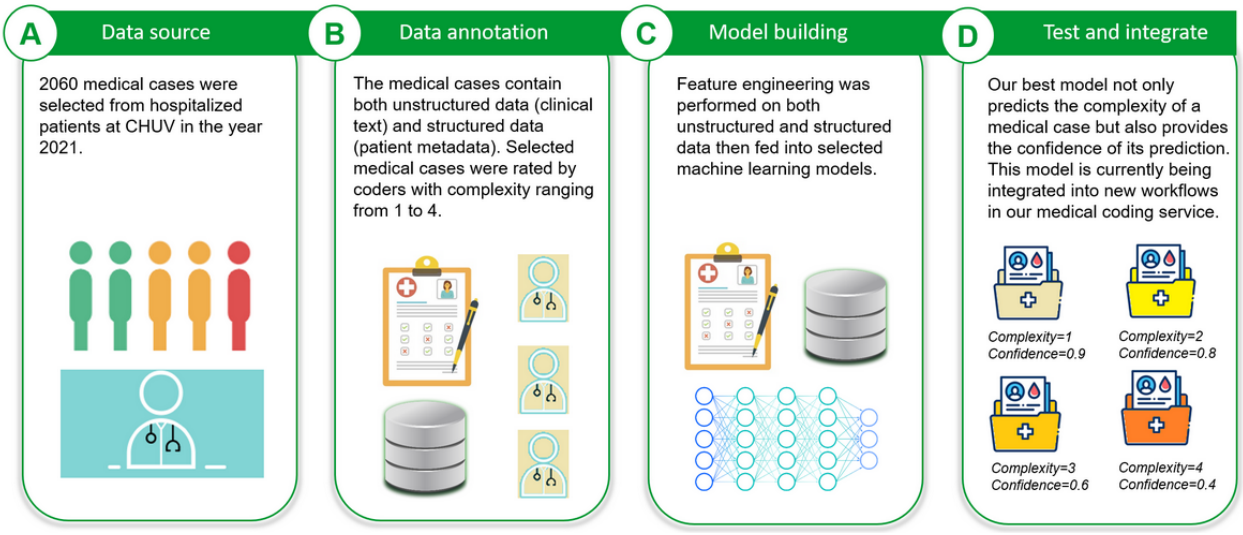
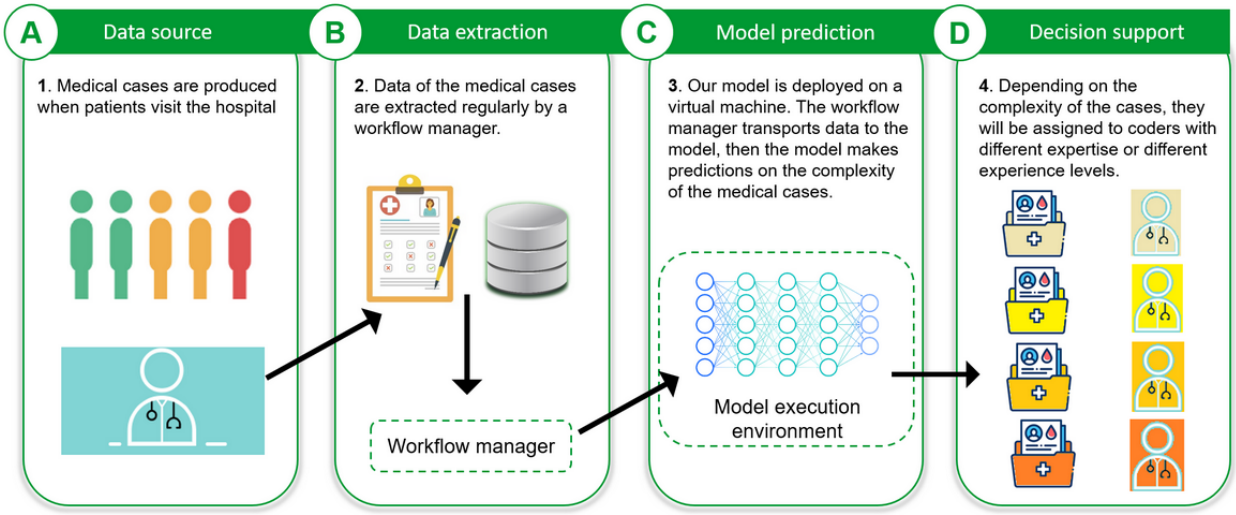


Figure 3. Integration of our model into the coding service. (A) When an inpatient visits the hospital and their medical case has been produced, the clinical text and patient metadata are stored in our clinical data warehouse. (B) A workflow manager will extract new medical cases regularly and send the data to our model. (C) Our model is containerized and deployed to an execution environment, where it performs the prediction for received cases. (D) Model predictions, together with the confidence of the predictions, are presented to the end users through a user interface to support task distribution in the coding service.



Definition of Complexity

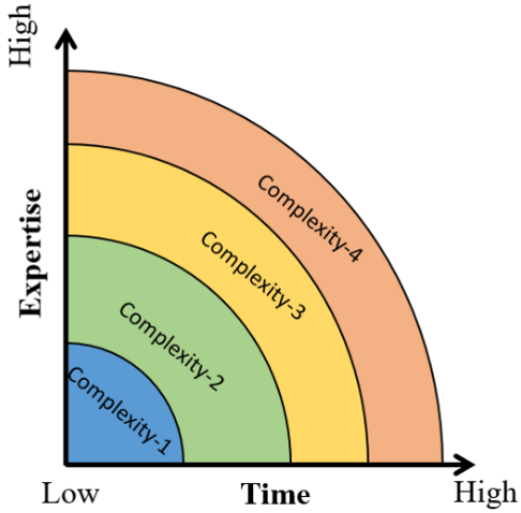
We use the term “coding complexity” to characterize the time and expertise required of medical coders to assign diagnostic codes to medical cases.

Expertise can be defined as the level of experience, medical knowledge, and mastery of coding rules. Therefore, a medical case can be complex by applying many coding rules without being difficult but increasing the possibility of attention errors. Other cases may be complex and difficult because of the medical knowledge they require for proper coding. Therefore, complexity was the measure chosen to categorize the cases.

If coding a medical case does not require much time and deep expertise, the coding complexity is low (level 1; Figure 4). Conversely, if coding a medical case requires a lot of time and deep expertise, the coding complexity is high (level 4; Figure 4).

Coding complexity, similar to pain or satisfaction, is a subjective quantity. A potential objective way of defining coding complexity can be provided by the automatic coding models. By passing the medical cases through automatic coding models and manually examining the confidence score and the completion and accuracy of ICD-10 code predictions, we could divide the cases into simple and complex groups. However, owing to the limited performance (ie, the very low recall score) of current automatic coding models regardless of language [12,20,21], this approach will not bring much value to our situation. Furthermore, if coding complexity could be measured using simple objective data (eg, similar to blood pressure), our multimodal modeling approach would be useless. Thus, in this study, our definition of coding complexity will focus on the subjective ratings provided by medical coders, aiming to minimize subjectivity by using ML approaches and to predict the subjective scores of complexity.

Figure 4. Intuitive representation of coding complexity regarding the time and expertise required of a coder.



To train our ML model, we extracted 2060 medical cases from hospitalized patients (inpatients) in 2021. We organized 2 annotation phases, each lasting 1 week, for 28 coders to rate the cases' complexity. During each annotation phase, the coders rated the complexity of the given cases based on an evaluation grid (Figure 4).

Data Collection and Preprocessing

Data Source and Data Annotation

A medical case contains 2 types of data: a patient's medical dossier and patient metadata (Textbox 1). We collected 2060 cases in total from the annotation phases. We note that the coding team at our hospital consisted of coders specialized in different medical domains. Hence, during annotation, we also kept track of whether a case was coded by a specialist. For example, if the responsible unit for a case was the internal medicine unit and the coder who coded this case was specialized in cardiology cases, the case was considered as not coded by its specialist coder.

Of the 2060 collected cases, 1998 (96.99%) were annotated by 28 medical coders, with each case coded by only 1 coder to maximize the size of the annotation set. As different medical coders may have different perceptions of the complexity of the same case, we evaluated the interrater reliability by asking 2 expert coders to code another 3.01% (62/2060) of cases. These 62 cases also represented our gold standard to create benchmarks for the models' performance. For case selection, we first trained several models using the 1998 cases; then used the best model's prediction to predict the complexity of several cases from our data warehouse; and, finally, randomly selected 62 out of the predicted cases while making sure that the complexity distribution of these 62 cases followed the same complexity distribution as the annotated data set. Each of the 62 cases was rated by each of the expert coders, and they were considered specialists for all cases. These 62 cases are referred to as the gold-standard set.

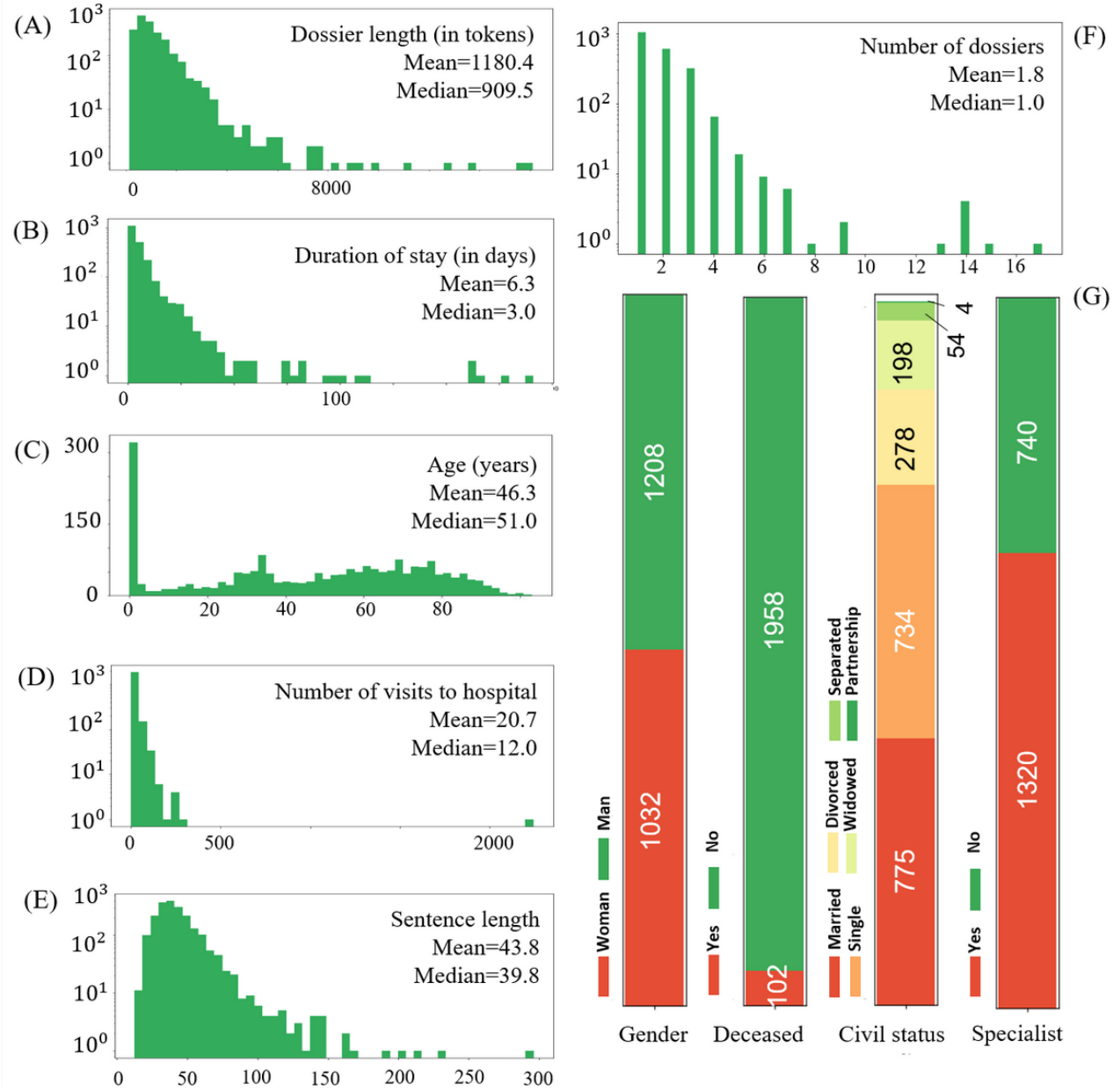
Textbox 1. Data collected for training and testing the model.

- Patient metadata: responsible medical service, number of movements between medical services, age, gender, civil status, whether the patient was deceased, length of stay, and whether the case was coded by a specialist
- Medical dossier: discharge letter of each service, operating procedure, intervention reports, and death letter

Metadata Preprocessing

The missing patients' metadata were imputed based on the nature of the data. For numerical values such as age and length of stay, the missing values were imputed with the median of the existing values because of their skewed distributions (Figure 5). For categorical values such as gender and civil status, the missing values were imputed with the mode of existing values.

Figure 5. An overview of the distribution of patient metadata per stay. Document length and sentence length are counted in terms of tokens (words and punctuation marks). The distributions on A, B, D, and E are heavily skewed. Note that the distributions on A, B, D, E, and F are log-scaled. The rightmost column of G is deduced from the coder's team specializations. The age=0 cases in C represent newborn cases.



Text Data Preprocessing

We tested both classic term frequency-inverse document frequency (TF-IDF)–based text encoding and ML-based text encoding, and different text preprocessing steps were applied accordingly. For TF-IDF text encoding, we first tokenized the text; then removed the stop words; and, finally, replaced the entities with their entity type. The second and third steps were used to reduce the noise and increase the frequency of important words to provide a better signal for the model. An example of processed text is presented in Textbox 2.

For ML-based text encoding such as fastText (Facebook AI Research lab) and transformers, no preprocessing was applied.

Textbox 2. An example of text preprocessing results.

- *Original text: Le patient susnommé a séjourné dans notre service du 01.02 au 03.02, date de son retour à domicile.*
- *Processed text: ["patient," "susnommé," "séjourné," "service," "<date>," "<date>," "date," "domicile," "."]*

Model Design

Overview

The overall approach of the model design was as follows. First, we extracted features from the preprocessed metadata and text data. Second, we tested 2 modeling approaches: framing the problem as a classification problem or as a regression problem. On the basis of the modeling approach, we used different metrics to evaluate the model performance.

Feature Engineering

As the values for the patients' metadata have different scales, we applied standardization (*z* score) to the numerical data and one-hot encoding to the categorical data.

To extract features from free text, we used 2 methods: TF-IDF and word embeddings.

TF-IDF provides a numerical weight of how important a word is to a collection of documents (Multimedia Appendix 1). We tested 2 configurations of the TF-IDF method: using the top 10,000 frequent terms or using the top 1000 frequent terms. We found that, using the top 10,000 frequent terms, the models performed better than using only the top 1000 frequent terms. Thus, in the following sections, we only report the results from the TF-IDF vector using the top 10,000 frequent terms.

Word embeddings provide the vectorized representation of a word based on the context in which it appears. We tested three types of word embeddings: (1) word2vec [22,23] embeddings trained on 2.5 million clinical texts (12 GB) collected from the hospital's clinical data warehouse; (2) the pooled output (CLS tokens) of the state-of-the-art French-language transformer model French-Language Understanding via Bidirectional Encoder Representations from Transformers (FlauBERT) [24], which was pretrained on 71 GB of French text collected from the internet; (3) the fastText supervised approach [25] with embeddings initialized with the pretrained word2vec embeddings of (1)—we tested fastText as it provided the subword approach that could reduce the impact of the out-of-vocabulary (OOV) issue. A detailed analysis of OOV for this study is provided in Multimedia Appendix 1.

Textbox 3 shows the sizes of the vectors extracted using the different methods. The detailed conversion methods are presented in Multimedia Appendix 1.

Textbox 3. Vector sizes of text feature engineering.

- Term frequency-inverse document frequency (vectors were extracted using scikit-learn [version 1.0.1]): 10,000
- fastText (initialized with customized embedding; fastText embeddings were extracted using fastText [version 0.9.2; Facebook Artificial Intelligence Research lab]): 100
- word2vec (customized; word2vec embeddings were trained using Gensim [version 4.0.0; RARE Technologies, Ltd]): 100
- French-Language Understanding via Bidirectional Encoder Representations from Transformers (FlauBERT; the FlauBERT embeddings and fine-tuned model were implemented using Hugging Face [version 4.17.0; Hugging Face, Inc]): 768

Model Architecture

The complexity of cases ranges from 1 to 4 with discrete values; thus, we can treat it as either a multi-class classification problem or as a regression problem. The tested models are presented in Figure 6.

For both classification and regression, we used different feature combinations as inputs to train the models. The combinations were as follows: (1) metadata only, (2) word embeddings only, (3) TF-IDF vectors only, and (4) TF-IDF concatenated with metadata.

The overall process of model implementation is summarized in Figure 7. During training, we applied 5-fold cross-validation to reduce overfitting. As the labels were unbalanced, we used stratified sampling for cross-validation in the classification models. We performed hyperparameter tuning of the most promising features and models. For TF-IDF, we optimized the number of words considered in the vocabulary (topmost frequent words) and text preprocessing (lower case, lemmatization, removal of stop words, and removal of nonalphanumeric tokens). For the gradient-boosted trees model, we tuned the number of estimators, learning rate, and maximum depth. Hyperparameters were tuned based on the average performance over all folds in the cross-validation sets using Bayesian optimization.

In addition, we tested the fine-tuning of the FlauBERT sequence classification model using the Hugging Face transformer library [26]. The FlaubertForSequenceClassification application programming interface provides a pretrained FlauBERT model with a classification layer of size 1024 on top. It takes raw text as input and outputs the predicted classes (in our case, which is the complexity level). Among all our experiments, our best results were obtained using the fine-tuned FlauBERT-base uncased model. Notably, we froze the first 11 encoder layers and trained the last encoder layer and the classification layer to limit overfitting. We also weighted each class differently in the cross-entropy loss to account for imbalance. We used the maximum sequence length of 512 tokens and a batch size of 32. In this manuscript, we only report the fine-tuned FlauBERT results obtained using this configuration.

Figure 6. Comparison of performance using different models and input features on the 5-fold–cross-validated training data set (1751 cases) and the best model performance on the test set (309 cases). Dashed vertical lines represent the baseline model results. Models are ranked based on the classification macro–F1-score in the figure. *Average per service: for a given case in a given service, it always predicts the average complexity of cases in this service. A total of 29 services have an average complexity of 2, a total of 5 services have an average complexity of 3, and a total of 1 service has an average complexity of 1. **Majority vote: always predicts the majority class (in our case, complexity 2) and serves as a baseline for model prediction performance. FlauBERT: French-Language Understanding via Bidirectional Encoder Representations from Transformers; TF-IDF: term frequency-inverse document frequency.

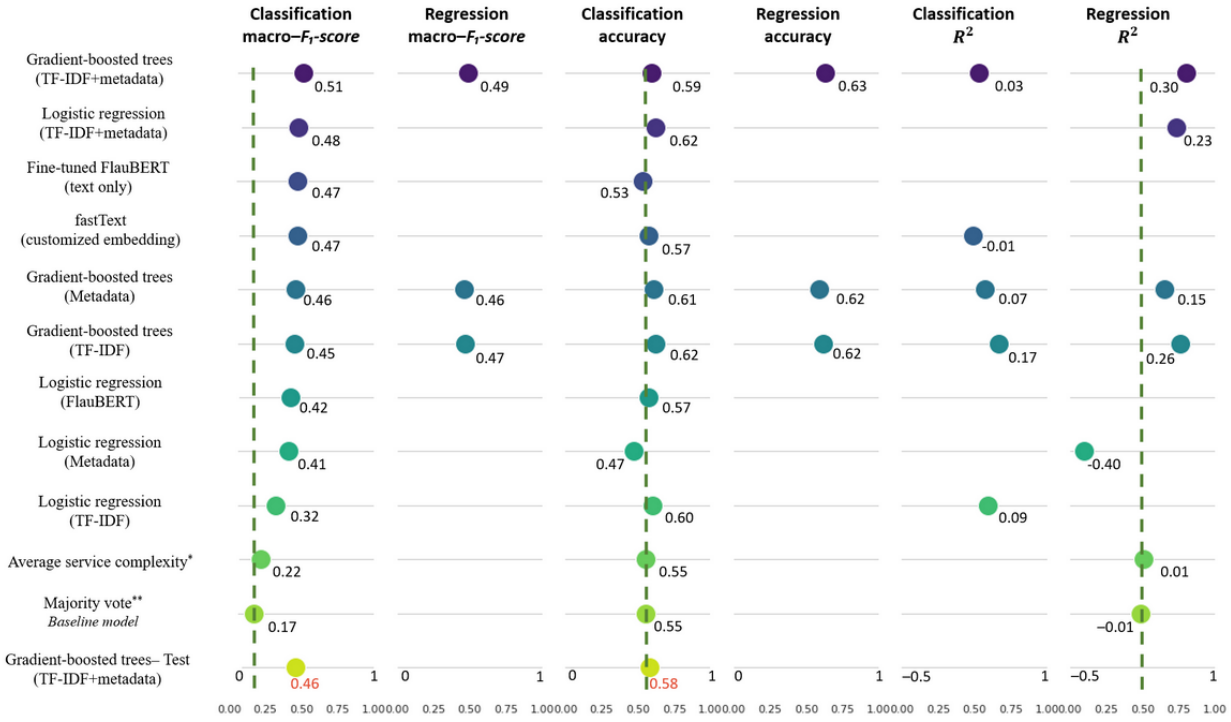
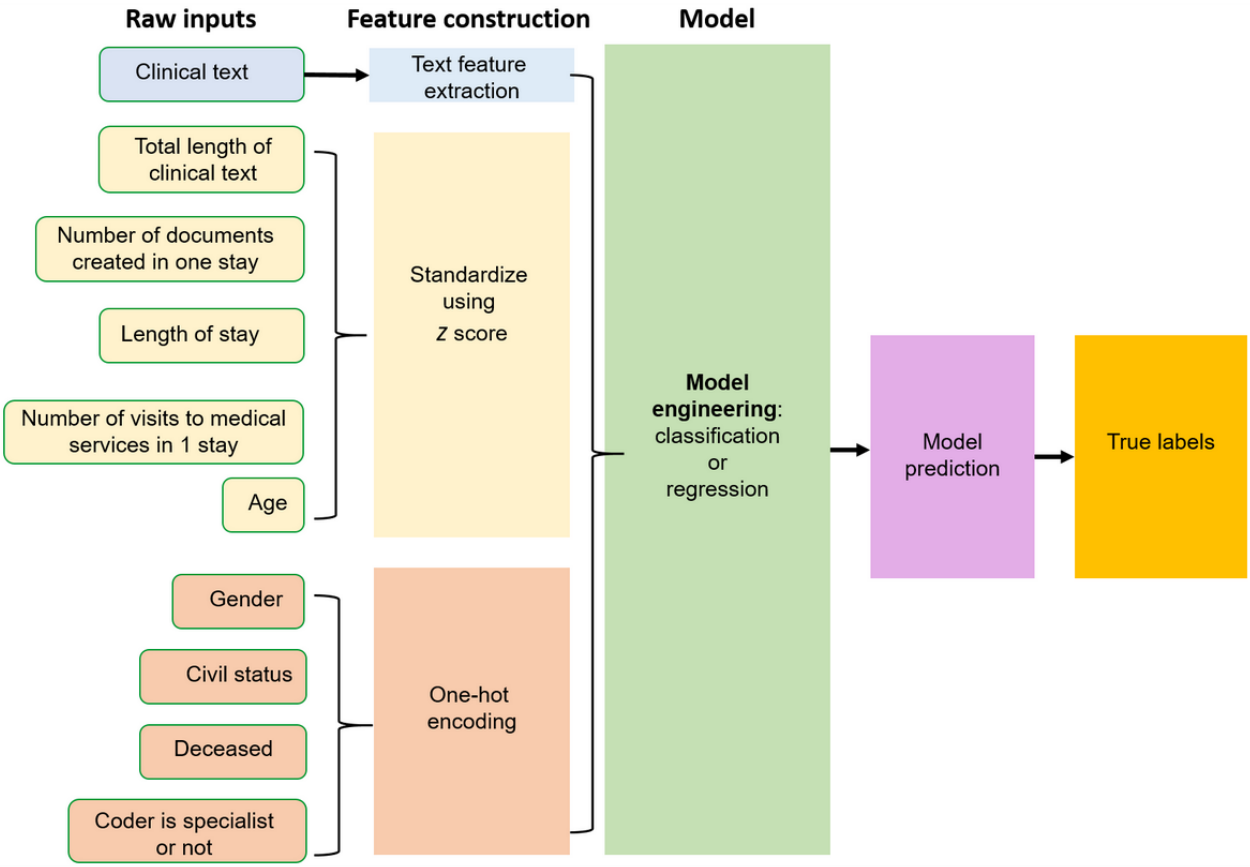


Figure 7. Feature engineering and modeling approach using word embeddings and patient metadata as model inputs. The fine-tuned French-Language Understanding via Bidirectional Encoder Representations from Transformers text classification model is not included in this flow.



Data Imbalance

Our data labels were strongly imbalanced, and we tried to overcome this issue by using oversampling and undersampling techniques. Our best model was trained using Synthetic Minority Oversampling Technique [27] for oversampling underrepresented classes followed by random undersampling for overrepresented classes. We also chose metrics to penalize models that did not predict underrepresented classes, such as the macro–F_1-score. Ordinal classification can also be an interesting "hybrid" approach. However, we leave trying more sophisticated classification approaches for future work.

Technological Stack

The ML pipeline leverages spaCy (version 3.1; Explosion AI) for preprocessing texts (using the French-language model "fr_core_news_md"), scikit-learn (version 1.0.1) to build complex pipelines that can work with cross-validation, and Optuna (version 2.10.0; Preferred Networks, Inc) to conduct hyperparameter searches. It also eases the deployment of the selected model as preprocessing is part of a single serialized pipeline. The other tools used to try other approaches were fastText for document classification, Gensim (RARE Technologies, Ltd) to manipulate pretrained word embeddings, and Hugging Face Transformers (Hugging Face, Inc) to use pretrained transformer models. Training was performed on a virtual machine with 64 central processing unit cores, allowing us to parallelize training, and an Nvidia RTX 3090 graphics processing unit for larger deep learning models.

The first version of the selected model is being deployed with Machine Learning Model Operationalization Management infrastructure in our medical coding service. The deployment details are presented in Multimedia Appendix 1.

Results

Metadata Analysis

Each team of coders had a set of medical specialties. We considered that a case was annotated by a specialist if the annotator was part of a team from one of the specialties involved in the case. Following this logic, 63.98% (1318/2060) of the cases were annotated by a specialist. We used this as a feature during training. At inference time, we could choose to request a prediction for whether the case would be coded by a specialist.

The distribution of the numerical metadata and categorical metadata is presented in Figure 5. To check if any of the metadata had significant predictive power on coding complexity, we performed Pearson correlations between the numerical metadata features and the complexity ratings; we also performed statistical tests on categorical features such as patient gender and marital status (Table 1). The results show that, in the precoding phase, features such as sentence length and number of medical services visited during a stay did not have strong effects on coding complexity. In the postcoding phase, the number of ICD-10 codes and Swiss Classification of Surgical Procedures codes showed correlations with coding complexity. With these results, we propose that a future direction of NLP-

or AI-assisted coding could use the metadata and clinical text to predict the number of codes that a case may produce and then compare it with the actual codes obtained after the coding process to perform quality checks in the postcoding phase.

Table 1. Pearson correlations between the numerical metadata features and the complexity ratings in both the pre- and postcoding phases and statistical tests of the categorical features and complexity ratings in the precoding phase.

	Correlation or statistical test	*P* value
Numerical features		
Number of tokens from all documents in a stay	0.44	<.001
Number of documents produced in a stay	0.33	<.001
Number of medical services visited during a stay	0.02	.35
Duration of the stay	0.41	<.001
Age	0.25	<.001
Sentence length	0.003	.83
Categorical features		
Marital status	$F_{5, 2054}=14.05$	<.001
Gender	$t_{2058}=-3.70$	<.001
Other metadata available after coding		
Number of ICD-10[a] codes	0.55	<.001
Number of CHOP[b] codes	0.46	<.001
DRG[c] cost	0.34	<.001

[a]ICD-10: International Statistical Classification of Diseases and Related Health Problems, 10th Revision.

[b]CHOP: Swiss Classification of Surgical Procedures.

[c]DRG: diagnosis-related group.

Coder Rating Analysis

The complexity ratings of the cases are shown in Figure 8A. The most common rating was complexity 2 (1127/2060, 54.71% of cases), and the least common rating was complexity 4 (58/2060, 2.82% of cases). We used stratified sampling to select the training and test sets; hence, their distributions were nearly identical to the true distribution shown in Figure 8A.

The original medical service of a case may also affect its complexity. Figure 8B shows that the cases from the Department of Palliative Care have the highest average complexity, whereas cases from the Department of Thoracic Surgery have the lowest average complexity.

By analyzing the gold-standard set, where all cases were rated by 2 experts, we found that even the expert coders did not always agree with each other. Of the 62 cases, the 2 experts agreed on 41 (66%). However, they disagreed by more than one complexity level in only 3% (2/62) of cases (Table 2). The interrater reliability (Cohen κ score) was 0.49 between the 2 expert coders. If we consider one expert as the ground truth and the other expert as a predictive model, the macro–F_1-score of this "predictive model" can only achieve 0.67 (Figure 9), a moderately good score showing that the task can be learned but models will not achieve a very high performance.

The reason why coders rate the same case with different complexity levels is mainly subjectivity. This is also a reminder that subjective-rated labels are often noisy, and no model can achieve a perfect performance. The ratio of agreement between 2 expert coders gives us an idea of the performance we could expect from a model. If we consider one expert as the model that predicts complexities and the other expert gives true complexity labels, then the highest accuracy that this model (the former expert) can achieve is 66%. In this sense, when later analyzing our model's performance, the 66% accuracy can be considered as one of the benchmarks. However, given the strong imbalance in the complexity labels, we should rely as well on the confusion matrix to compare the annotator-annotator agreement with the model-annotator agreement.

However, as mentioned in the Model Design section, our samples were highly imbalanced, and the accuracy metric lacked the ability to measure the model's performance comprehensively according to the sample distribution. As there were 54.71% (1127/2060) of cases rated with a complexity of 2, a naive model that predicts 2 all the time could reach an accuracy of 54.71%, but it provides no value for solving our problem. To consider the imbalanced sample distribution, we used the macro–F_1-score together with accuracy to measure the model performance. The macro–F_1-score between the 2 coders was 0.67, which was considered as the other benchmark that we used to evaluate the model's performance.

Figure 8. (A) The distribution of complexity ratings over all 2060 cases. (B) Average complexity rating by service. The green bars show the top 5 services, and the red bars show the bottom 5 services. CHT: thoracic surgery; ION: immuno-oncology; MIN: infectious diseases; OBS: obstetrics; PED: pediatrics; RHU: rheumatology; SIA: adult intensive care; SIP: pediatric intensive care; SPL: palliative care; URG: emergency department.

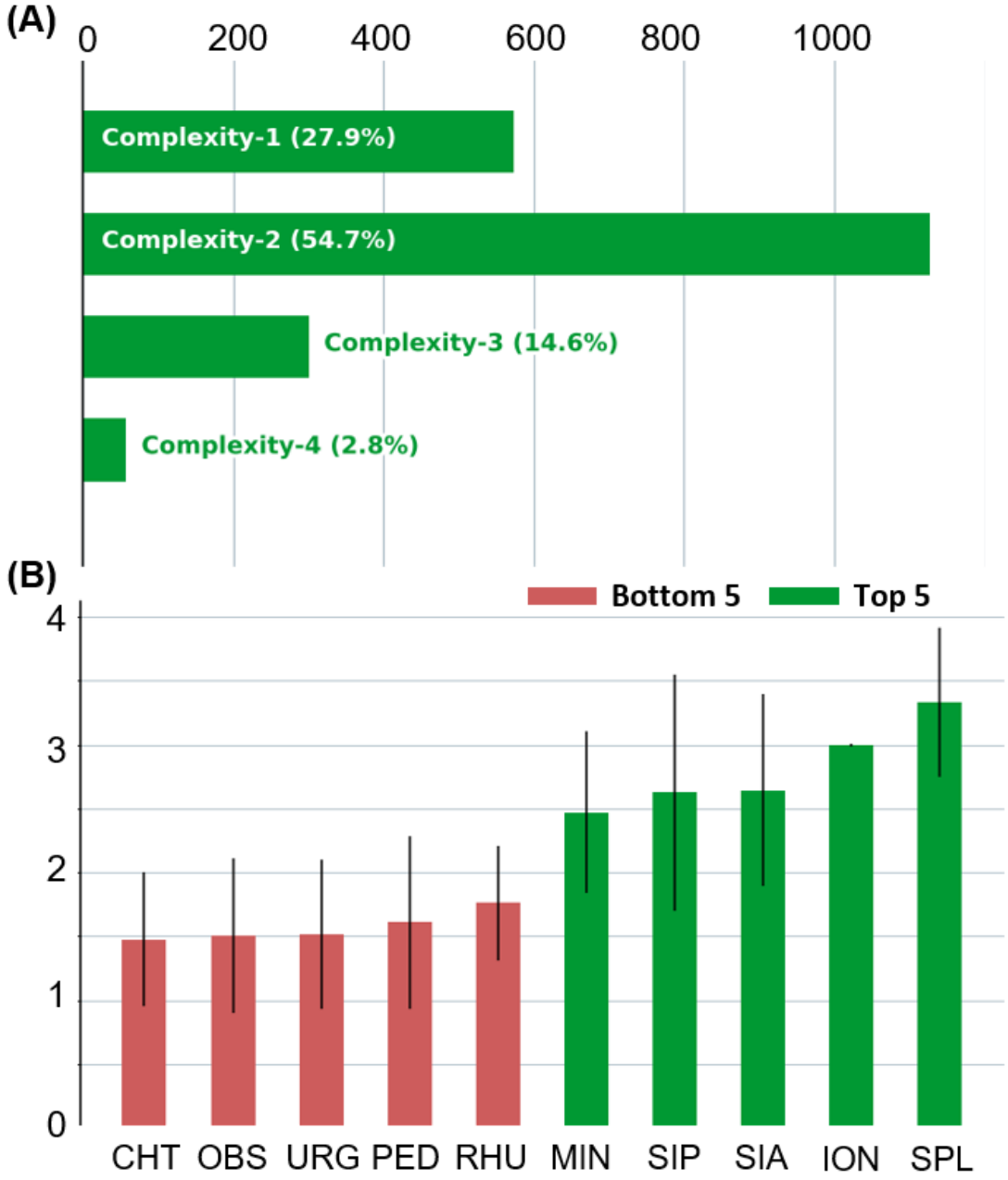


Figure 9. (A) The complexity rating comparison between 2 expert coders on the gold-standard set. (B) The comparison between the validation model's predictions and average expert ratings on the gold-standard set. (C) The comparison between 2 expert coders' ratings on the gold-standard set when grouping into simple (complexity 1 and 2) and complex (complexity 3 and 4) cases. (D) The comparison between average expert ratings and the validation model's predictions on the gold-standard set when grouping into simple and complex cases. The average expert ratings are rounded up to the next largest integer.

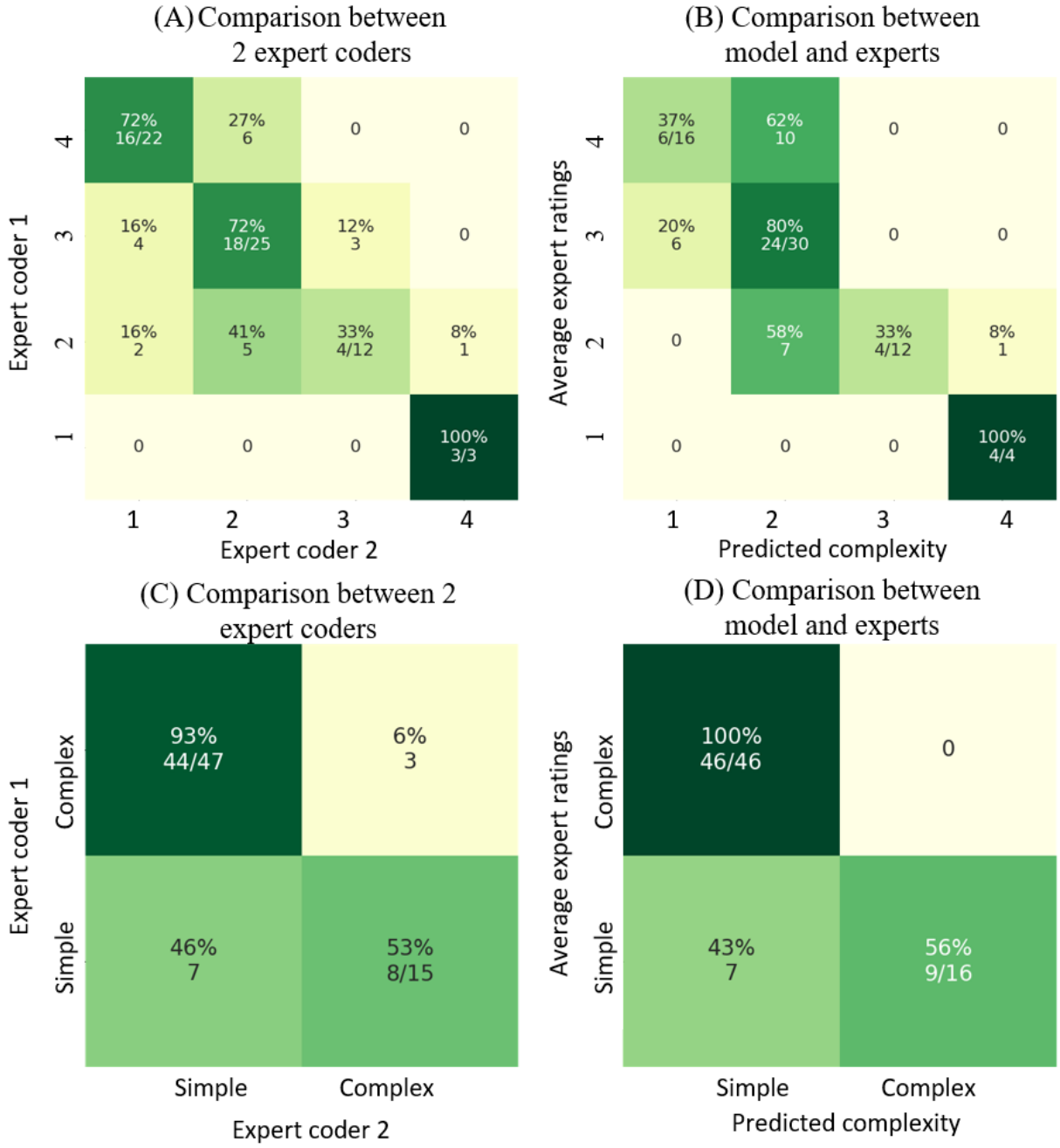


Table 2. Absolute difference between expert 1 and expert 2 complexity ratings. The accuracy reached by expert coders was approximately 66% (41/62; N=62).

Absolute difference in complexity ratings between expert coders 1 and 2 (number of complexity levels)	Cases, n (%)
0	41 (66)
1	19 (31)
2	2 (3)
3	0 (0)

Model Analysis

Overview

First, we wanted to study whether our approach worked on predicting coding complexity for medical cases. We made use of all the 2060 annotated cases (n=1998, 96.99% 1-coder–rated and n=62, 3.01% gold-standard cases). We split the 2060 cases into a training set (n=1751, 85% of cases) and a test set (n=309, 15% of cases) and tested our model architecture. Then, to validate the model's performance with expert coders' benchmarks, we left the 3.01% (62/2060) of gold-standard cases out as the test set and trained a model with the same architecture but with more training data (1998/2060, 96.99% of cases).

The Main Model

To train the models, we started by using either patient metadata only or word embeddings or TF-IDF vectors only as input features. The best-performing model using patient metadata was gradient-boosted trees (macro−F_1-score=0.46; accuracy=0.61 for classification; R^2=0.15 for regression). The best-performing model using word embeddings was the fastText classification model (macro−F_1-score=0.47; accuracy=0.57; initialized with customized embeddings), and the best-performing model using TF-IDF vectors was gradient-boosted trees (macro−F_1-score=0.45; accuracy=0.62 for classification; R^2=0.26 for regression).

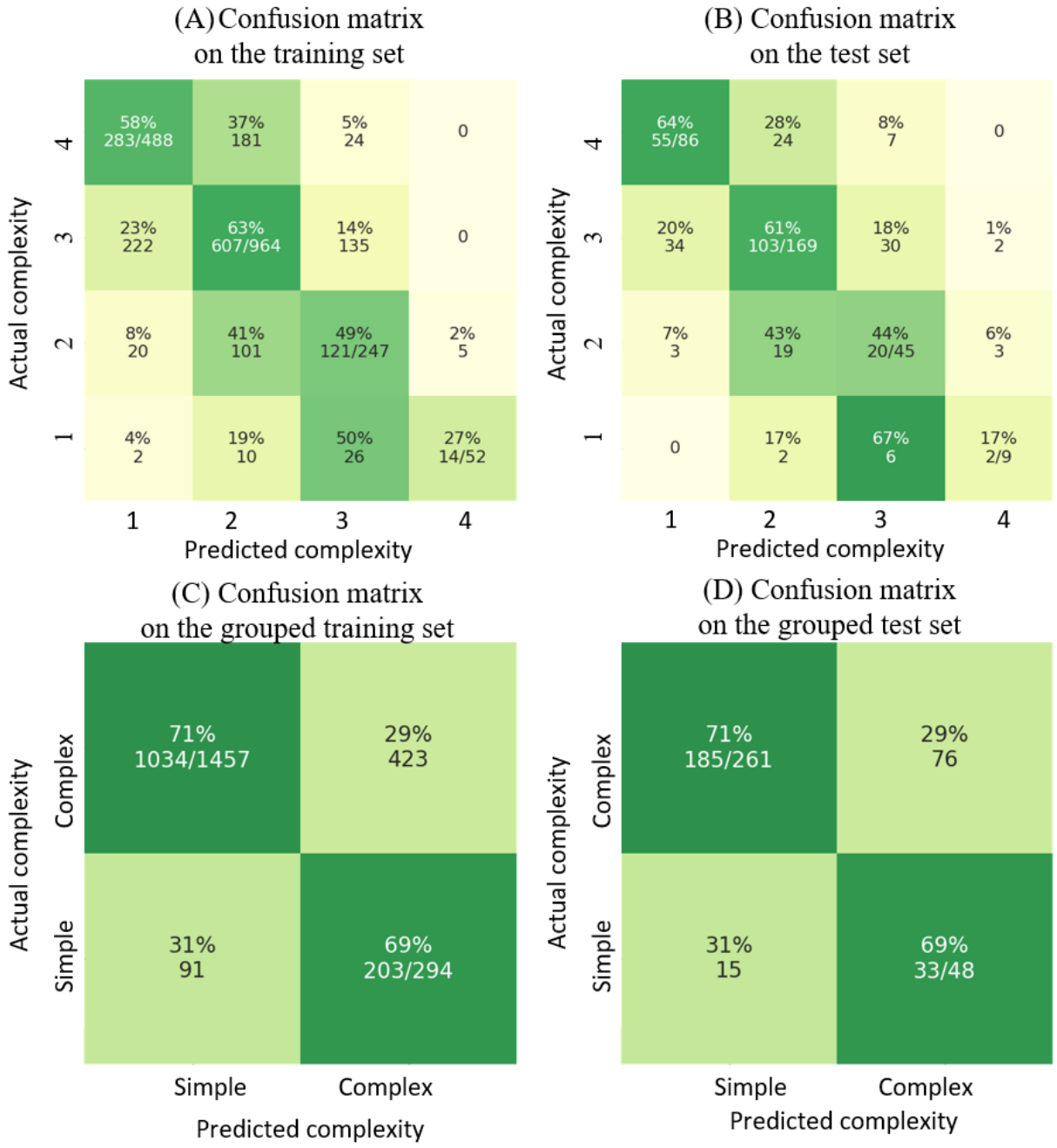
The model using word embeddings did not outperform the model using TF-IDF vectors. Thus, we combined the TF-IDF vectors with metadata as input features to integrate information from both patient metadata and medical dossiers. The best-performing model used gradient-boosted trees and achieved a macro−F_1-score of 0.51 and accuracy of 0.59 on the cross-validated training set and a macro−F_1-score of 0.46 and accuracy of 0.58 on the test set. Figure 6 shows the performance comparison between different models on the 5-fold–cross-validated training data set and the test set. The detailed numbers can be found in Multimedia Appendix 1.

As performing well on underrepresented classes is important in our case, we report the macro–F_1-score as the first metric. Macro–F_1-score is the average of the F_1-score per class and is not weighted by the number of instances in the class. Unlike accuracy, this metric penalizes each class equally. On the basis of the macro–F_1-score, we selected our best model as the gradient-boosted trees trained with the combined TF-IDF and metadata features (referred to as the main model).

The confusion matrix (Figures 10A and 10B) shows that our main model confused complexity-2 and complexity-3 cases during training and testing. Figure 9A shows that, even for expert coders, there was no clear distinction when rating complexity 2 and 3 for a case. The difficulty to distinguish between complexity 2 and 3 could be due to the similarity between the 2 classes of cases. We noticed that our main model also had difficulties distinguishing between complexity 3 and 4 during training and testing. This performance could be due to the lack of examples. Although we performed oversampling using Synthetic Minority Oversampling Technique on cases with a complexity of 3 and 4, it still lacked variability in complexity-4 cases.

We then tried to merge complexity-1 and complexity-2 cases as "simple" cases and complexity-3 and complexity-4 cases as "complex" cases and tested the model as a binary classifier. The results (Figures 10C and 10D) show that the model performed well on distinguishing between simple and complex cases. On the training set, the model achieved a macro–F_1-score of 0.62 with an accuracy of 0.71. On the test set, the model achieved a macro–F_1-score of 0.65 with an accuracy of 0.71.

Figure 10. (A) and (B) The main model's performance on the training set (1751 cases) and the test set (309 cases). (C) and (D) The main model's performance on the grouped training set (1457 cases as simple and 294 cases as complex) and the test set (261 cases as simple and 48 cases as complex).



The Validation Model

To validate our model approach and compare it with experts' benchmarks, we trained a validation model using the 96.99% (1998/2060) of 1-coder–rated cases and tested it on the 3.01% (62/2060) of gold-standard cases. The architecture of the validation model was the same as that of the main model.

The comparison between the 2 expert coders' ratings (Figure 9A) shows that most of the expert coders' disagreements were on complexity-2 and complexity-3 cases, and the overall agreement ratio between the 2 coders was 66% (41/62), with a macro–F_1-score of 0.67. Table 3 and Figure 9B show the comparison between our validation model and the 2 experts' ratings on the gold-standard set. The model agreed on 53% (33/62) of the cases with expert coder 1 and in 63% (39/62) of the cases with expert coder 2. The validation model achieved a 61% agreement ratio with the average ratings of both experts, with a macro–F_1-score of 0.62.

Table 3. Comparison between our validation model's predictions and 2 expert coders' ratings on the gold-standard set.

	Percentage of agreement	Pearson correlation
Expert coder 1 vs expert coder 2	66	0.70[a]
Model vs expert coder 1	53	N/A[b]
Model vs expert coder 2	63	N/A
Model vs ceiled mean of 2 expert coders	61	0.70[a]

[a]$P<.001$.

[b]N/A: not applicable.

When merging the 4 complexity levels into 2 (simple vs complex; Figures 10C and 10D), the agreement ratio between the 2 coders became 84% (52/62) with a macro–F_1-score of 0.76, and the agreement ratio between model predictions and average expert ratings became 0.89 with a macro–F_1-score of 0.82. The results indicate that the model is comparable with human experts' performance and predicts in a very similar manner to that of human experts (Figures 9A and 9B).

Interestingly, for the gold-standard cases, our validation model managed to predict complexity-4 cases 100% correctly, which was different from the main model's performance during training and testing (Figures 10A and 10B). As there were only 4 selected cases with a complexity of 4 owing to the sampling for expert cases, these cases could be extremely complex and, thus, easy for the model to identify.

Compared with other models that can provide higher accuracy but lower F_1-score, both the main model and the validation model were more practical in our concrete use case as it is important to predict diverse complexity levels rather than keep predicting a complexity of 2 for all cases (Multimedia Appendix 1).

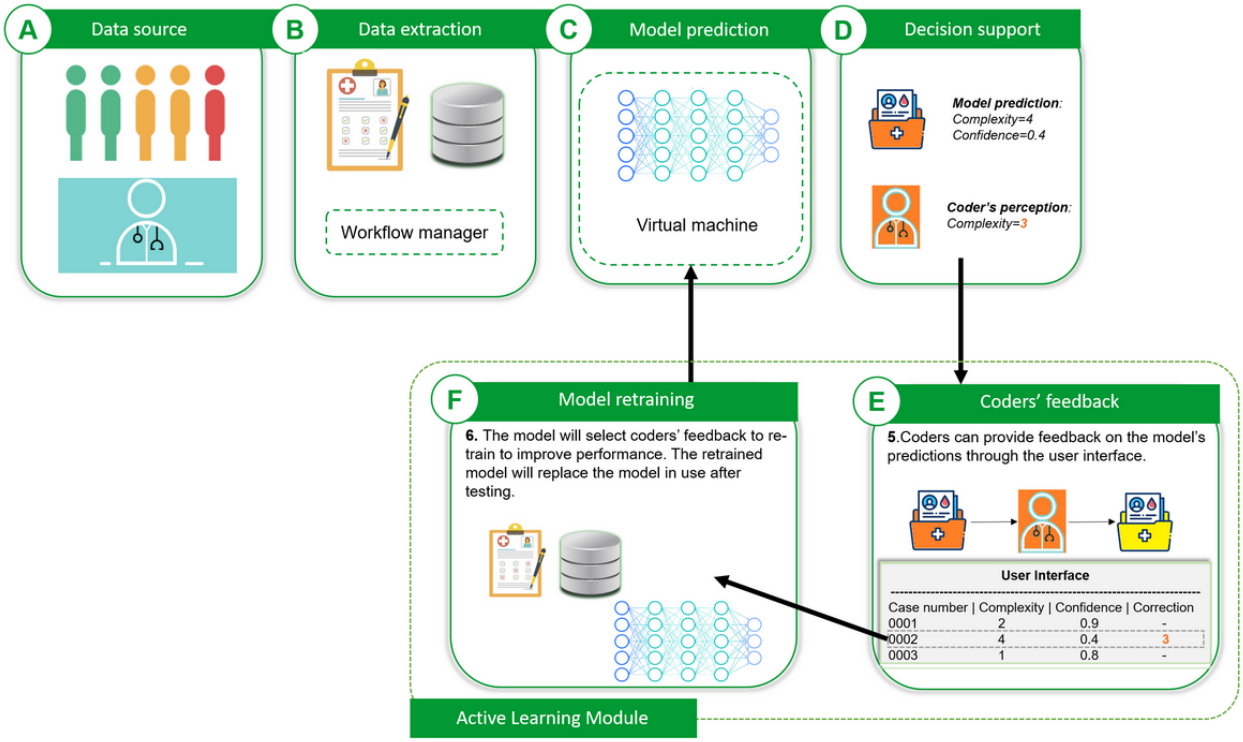
Classification Versus Regression

We summarize the pros and cons of both approaches given our use case in Textbox 4.

Textbox 4. Pros and cons of the classification and regression approaches.

- *Prediction confidence*: many classification models output the confidence in the predicted class as a probability, whereas regression models typically do not provide such information out of the box (although CIs are sometimes possible). Confidence is useful for end users, meaning that they can disregard predictions with low confidence. It can also be used in the active learning module (Figure 11) to select new cases (with low prediction confidence and strong disagreement between prediction and coder perception) to retrain the model.
- *Interpretability of results*: using a classification approach enables the computation of F_1-scores, accuracy, and confusion matrices. These are more intuitive for end users. Note that, for regression, it is still possible to round prediction to apply these metrics.
- *Order of labels*: complexity scores are naturally ordered. Therefore, given a case annotated with a complexity of 4, a model should be penalized more for predicting a complexity of 1 than for predicting a complexity of 3. Regression methods consider order, whereas classification methods do not.

Figure 11. Use of active learning module to collect coders' feedback and improve model performance. The workflow manager in (B) can be any software or platform that provides automatic scheduling for designated work (eg, a script for data extraction).



Discussion

Principal Findings

We presented different ML models that can predict the complexity of coding medical cases with 4 complexity levels. We first trained the models on all 2060 annotated cases. When only using patient metadata, the best model (gradient-boosted trees) could achieve a macro−F_1-score of 0.46, an accuracy of 0.61 for classification, and an R^2 of 0.15 for regression. By applying NLP methods to extract information from clinical text, the best model (fastText initialized with customized embeddings) could achieve a macro−F_1-score of 0.47 and an accuracy of 0.57 for classification. When combining patient metadata and NLP-extracted information, the best model (the main model in the Model Analysis section) achieved a macro−F_1-score of 0.51 and an accuracy of 0.59 on the cross-validated training set and a macro−F_1-score of 0.46 and an accuracy of 0.58 on the test set.

To evaluate our model approach with experts' benchmarks, we trained our validation model using the same architecture as the main model on all except the gold-standard cases. Our validation model achieved an accuracy of 0.61 with a macro−F_1-score of 0.62 on the gold-standard cases. When merging the 4 complexity levels into "simple" (complexity 1-2) and "complex" (complexity 3-4) cases, our validation model could achieve an accuracy of 0.89 and a macro−F_1-score of 0.82. The results indicate that the model performance is highly comparable with that of human experts.

To the best of our knowledge, this is the first study to apply NLP and ML models to help differentiate the complexity of coding medical cases.

Clinical Importance

Lausanne University Hospital in Switzerland has 2 missions: guaranteeing medical services in an area and serving as a referral hospital. The dominance of cases with a complexity level of 2 (referred to as case 2) in the labeled sample cases can be explained by this double activity as the hospital not only concentrates on university or referred complex cases but also receives normal cases similar to other hospitals.

In our current medical coding service, the cases to be coded are distributed 50% to the team of the specialty and 50% to a "common pot." This team versus common pot distribution is done randomly without considering the complexity of the cases, leaving complex cases in the common pot and, conversely, depriving the common pot of "simple" cases of specialized resources. Note that, in our case, coders can still choose complex cases from the common pot even if the case is not in their specialty. Many coders care about diversity or learning other types of cases. The integration of this model enables them to choose the complexity consciously.

The dominance of cases 2 will have the effect of pushing a lot of cases into the common pot, reducing the number of cases arriving to teams of different specialties and, hence, reducing the ratio of common pot to specialists. The quality of coding of complexity-3 and complexity-4 cases will be improved as they will be redirected to the specialty teams or senior coders. However, this will also be at the risk of lowering the quality of coding of cases 2, which will end up in the common pot. Therefore, it will be necessary to maintain a 50/50 ratio between

the common pot and the teams or senior coders and force cases 2 to be coded by teams or seniors as well. This adjustment will enhance the quality of coding of cases 3 and 4 and a maximum of cases 2. After our system is deployed, the new distribution considering the complexity predicted by our NLP and ML model will be monitored in terms of satisfaction of the coding teams and accuracy of coding. Furthermore, we will analyze the accuracy of coding in relation to the predicted case complexity to adjust the model design and more efficiently allocate the case distribution to coders.

In our current model, the complexity of the cases is defined by the coders from our medical service and is rated subjectively. By analyzing the model predictions for a variety of cases, it is possible to summarize the common features shared by the high-complexity cases and those shared by the low-complexity cases. The summarized features can be used to build a set of objective rules that can be shared with other clinical services or the medical coding services of other hospitals. For small hospitals or clinical services, which do not always have sufficient resources to train and build their own ML models, this set of rules can help them distribute the cases more efficiently. In contrast, if the summarized features could not distinguish well between the simple and complex cases, it may reflect that the case complexity is a subjective rather than objective measure. In this situation, the best way to generalize this subjective measure is to build a model, such as in our approach, to learn the highly nonlinear subjective measures.

The complexity of coding a medical case can approximately reflect the complexity of the corresponding clinical case. Our application can not only improve resource allocation in medical coding services but also be generalized to other clinical services. Indeed, coding complexity levels can also be used in decision-making processes to help arbitrate resource allocation among professionals in the same department but affiliated with different clinical services within the department. For example, in the surgery department, a similar approach can be applied to help study the need for resources for different subspecialties based on the volume of treated cases but also on their relative complexity. The generalized application can be integrated into different digital health care systems for automatic task assignment to avoid conflicts in an unfair workload distribution.

Technical Importance

OOV is an issue that can impair model performance. Although the word2vec embeddings used in this study were trained on our own clinical data, OOV was still present as the corpus we used to train the embeddings might not have been sufficient to cover all the clinical terms used in the medical discharge documentation. To mitigate the impact of OOV, we tested the fastText subword approach. However, as shown in the Model Analysis section, the model performance was not much improved because of the low OOV ratio of our data set, which was only approximately 8% in the 2060 selected cases for this study. We provide a detailed analysis of OOV in our corpus in Multimedia Appendix 1.

As new clinical documents are produced every day, our deployed model could also face the impaired performance caused by the OOV issue. The solution we propose in this paper to reduce the impact is to monitor the evolution of new OOV with respect to the training data set and retrain the word embeddings when needed. During the retraining phase, we will not only retrain the word embeddings but also retrain the models with coder feedback to further improve the model performance from the perspective of both feature engineering and model engineering.

In our study, we used FlauBERT, which is a pretrained French-language transformer, in 2 different ways. The first way to use it is to generate word embeddings as text features for model inputs. We then also tested a Hugging Face [26] implementation of the sequence classification model using FlauBERT. A detailed description of this approach is presented in Multimedia Appendix 1. The best performance using the transformer model directly achieved a macro–F_1-score of 0.47, which is similar to other models that only receive text as features. The model performance did not improve as much as expected. The reason could be that our data set was too small (only 2060 cases) compared with the size of the transformer model. Regarding this, we will continue collecting coder feedback on the predicted cases and use them to train the model continuously. With these approaches, we hope to improve the transformer model performance in the future.

We found that using TF-IDF vectors as text features provided better prediction performance than using word embeddings as text features. The fastText and FlauBERT embeddings were pretrained on a nonclinical corpus; thus, the represented context of the word could deviate from the context used in the clinical text. As shown in the Metadata Analysis section, the median document length per stay was 909 tokens. Common pretrained transformer-based models handle up to 512 tokens, and it is not obvious which subset of the document should be selected to pass to the model. Although it is possible to overcome this limitation by embedding each chunk of 512 tokens and averaging their embeddings, we believe that a substantial improvement over other methods is needed to justify the computation cost. Furthermore, fastText and word embeddings both perform averaging over all vectors of each document, which may dilute the signal too much given the number of tokens. In contrast, TF-IDF can preserve some of this information, which could be the reason why TF-IDF vectors outperformed word embeddings in our task. A future direction to improve the model performance could be to combine TF-IDF vectors with word embeddings as text features. TF-IDF vectors can be used as a weight of importance for the words, whereas word embeddings can represent the contexts of the words. By combining the two, we could obtain vectors that represent both the importance and context of the words comprehensively. Another possible approach to improve the model performance is to build a rule-based model from coders' experiences and then combine the rule-based model with the ML model, which can increase both the interpretability and flexibility of the prediction. As the complex cases are more likely to have multiple laboratory tests and clinical examinations, we could also include this structured clinical information for future feature engineering.

By comparing our model's predictions with the expert coders' ratings, we found that the model could achieve an expert performance level (Figure 9). As rating case complexity is relatively subjective, even expert coders do not always agree with each other. This introduced another level of complexity to our study. However, by learning 1998 cases from the training set, our model's performance became comparable with that of the experts.

One of the advantages of our model is that we used a multimodal approach. Structured data such as patient metadata can provide quantitative information about patients' status. Clinical text can provide rich information on diagnostic and other assessments of patients, which are not usually presented in the structured data. By combining the two, we are able to maximize the information needed to evaluate the complexity of a clinical case. Our study used 1 model to process data of different modalities and make predictions. In future work, we propose using dedicated models for each data modality and combining the predictions of multiple models using another ML model to make the final prediction. The benefits of using multiple models are that (1) it is easy to plug in new data and new models into the architecture, which makes the model flexible to extend, and (2) it is easier to perform feature engineering and interpret the model's prediction.

The advantage of classification models over regression models in our study was that classification models allowed us to produce the confidence of the predictions. By showing both the predicted complexity level and the confidence of the prediction, we are able to provide comprehensive information to end users. However, there are also limitations to our model. Of the 2060 cases we collected for this project, 54.71% (1127/2060) were labeled as complexity-2, and only 2.82% (58/2060) were labeled as complexity-4. The unbalanced data set affects the performance of the classification models, meaning that the models have a higher tendency to predict complexity 2 for a given case. This problem was tackled by oversampling the underrepresented cases and undersampling the overrepresented cases. The results showed that the model performed better with oversampling and undersampling techniques (Multimedia Appendix 1).

Our model will be integrated into our current coding system with an active learning module. Figure 11 shows the integration architecture. The model reads patient metadata and medical dossiers regularly from our clinical data warehouse through a workflow manager. The predictions are presented in the user interface of the coding software. When coders find that the prediction deviates from the perceived complexity, they can put their corrections in a feedback field. Coders' feedback is stored and sent to the model for retraining. This integration architecture allows us to track and continuously improve the performance of the model.

Future Work

Future work can be carried out on different aspects. To improve the model prediction performance, we can continue working on feature and model engineering. In addition to the data we used in this study, there could be other patient data that can be useful to predict the complexity of cases. Regarding the text features, we could try different combinations of NLP tools to maximize the information extraction from clinical text. We will also continue working on reducing the OOV impact by retraining the word embeddings (both word2vec and fastText) and TF-IDF vectors every 6 months and use coder feedback as new training samples to retrain the models. To make full use of the advanced transformer models, we will not only keep training using the new samples but also explore ways to incorporate patient metadata into the model design. We will also work together with coders to establish a sound and interpretable rule-based model and then combine it with the ML model. The hybrid model can provide both flexibility and good reasoning in distinguishing cases.

Currently, most NLP applications focus on AI-assisted coding using rule-based or ML models. As stated before, the rules framing medical coding complexity are dynamic and change over time, preventing the rapid learning of the tool. Instead of using AI-assisted tools only for coding, it is possible to extend the AI-assisted scope from case preselection to postcoding quality checks. Our approach provides a possibility to preselect cases that are suitable for automatic coding and other cases for manual coding. After a case is coded, AI-assisted tools can provide a post hoc analysis of the code categories and combinations, aiming to find possible mistakes in the codes. This can be done by studying previous coded cases using statistical and NLP analysis.

We also aim to continuously evaluate the application's impact on our medical coding service. After the integration, we will monitor the average time a coder spends coding a case and the average number of mistakes a coder makes for each case. By comparing the time and accuracy before and after the integration, we can obtain a quantitative measure of how much improvement the model can bring to the coders' daily work.

In addition to monitoring the quality of coding, we will keep tracking the coders' user experience. With the help of the active learning module, we are able to collect coders' feedback on the model's predictions. The model will be retrained based on coders' feedback through iterations to improve the prediction performance. As discussed in the Clinical Importance section, our application can not only help with task distribution to current coders but also be used to select cases for training junior coders. Junior coders will receive simple cases at the beginning and gradually receive more complex cases. This approach can give junior coders enough exposure to a variety of cases with respect to their capabilities as well as evoke their interests in medical coding.

Acknowledgments

The authors thank the 2 expert coders, Mireille Nya Buvelot and Lionel Comment, and all coders in the Coding Division for their contribution to complexity annotations. They also thank Dr Mostafa Ajalloeian for providing advice on this project.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Illustrations on the text feature engineering, imbalanced data processing, MLOps infrastructure, model comparison table, OOV analysis, and transformers fine-tune methods.
[DOCX File , 538 KB - medinform_v11i1e38150_app1.docx]

References

1. What is Medical Coding? American Academy of Professional Coders. 2021. URL: https://www.aapc.com/medical-coding/medical-coding.aspx [accessed 2022-03-14]
2. Iglehart JK. The new era of prospective payment for hospitals. New England Journal of Medicine 1982 Nov 11;307(20):1288-1292. [doi: 10.1056/nejm198211113072036]
3. Mayes R. The origins, development, and passage of Medicare's revolutionary prospective payment system. J Hist Med Allied Sci 2007 Jan;62(1):21-55. [doi: 10.1093/jhmas/jrj038] [Medline: 16467485]
4. Chilingerian J. Origins of DRGs in the United States: a technical, political and cultural story. In: Jimberly J, de Pouvourville G, d'Aunno T, editors. The Globalization of Managerial Innovation in Health Care. Cambridge, UK: Cambridge University Press; 2008:4-33.
5. International Statistical Classification of Diseases and Related Health Problems 10th Revision. World Health Organization. 2019. URL: https://icd.who.int/browse10/2019/en#/ [accessed 2022-03-14]
6. Roger France FH. Case mix use in 25 countries: a migration success but international comparisons failure. Int J Med Inform 2003 Jul;70(2-3):215-219. [doi: 10.1016/s1386-5056(03)00044-3] [Medline: 12909172]
7. Browne JH. High performance work strategies: empowerment or repression for the working class? J Bus Econ Res 2005 Jul 1;3(7):1-4. [doi: 10.19030/jber.v3i7.2788]
8. Johnson AE, Pollard TJ, Shen L, Lehman LW, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016 May 24;3:160035 [FREE Full text] [doi: 10.1038/sdata.2016.35] [Medline: 27219127]
9. Baumel T, Nassour-Kassis J, Cohen R, Elhadad M, Elhadad N. Multi-label classification of patient notes: case study on ICD code assignment. In: Proceedings of the Workshops at the 32nd AAAI Conference on Artificial Intelligence. 2018 Presented at: AAAI '18; February 2-7, 2018; New Orleans, LA, USA p. 409-416. [doi: 10.48550/arXiv.1709.09587]
10. Chen J, Teng F, Ma Z, Chen L, Huang L, Li X. A multi-channel convolutional neural network for ICD coding. In: Proceedings of the IEEE 14th International Conference on Intelligent Systems and Knowledge Engineering. 2019 Presented at: ISKE '19; November 14-16, 2019; Dalian, China p. 1178-1184. [doi: 10.1109/iske47853.2019.9170305]
11. Li M, Fei Z, Zeng M, Wu FX, Li Y, Pan Y, et al. Automated ICD-9 coding via a deep learning approach. EEE/ACM transactions on computational biology and bioinformatics 2019;16(4):1193-1202. [doi: 10.1109/TCBB.2018.2817488] [Medline: 29994157]
12. Kim BH, Ganapathi V. Read, attend, and code: pushing the limits of medical codes prediction from clinical notes by machines. In: Proceedings of Machine Learning for Healthcare Conference. 2021 Presented at: PMLR '21; August 6-7, 2021; Virtual p. 196-208 URL: https://proceedings.mlr.press/v149/kim21a/kim21a.pdf
13. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020 Feb 15;36(4):1234-1240 [FREE Full text] [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
14. Dalloux C, Claveau V, Cuggia M, Bouzillé G, Grabar N. Supervised learning for the ICD-10 coding of French clinical narratives. In: Proceedings of 2020 Medical Informatics Europe. 2020 Presented at: MIE '20; April 28-May 1, 2020; Geneva, Switzerland p. 1-5 URL: https://hal.archives-ouvertes.fr/hal-03020990/
15. Azam SS, Raju M, Pagidimarri V, Kasivajjala VC. Cascadenet: an LSTM based deep learning model for automated ICD-10 coding. In: Proceedings of the 2019 Future of Information and Communication Conference. 2019 Presented at: FICC '19; March 14-15, 2019; San Francisco, CA, USA p. 55-74. [doi: 10.1007/978-3-030-12385-7_6]
16. NLP/Forschung: Des traitements efficaces et rentables grâce à une technologie intelligente. ID Suisse AG. 2021. URL: https://www.id-suisse-ag.ch/fr/produits/nlp-forschung/ [accessed 2022-03-14]
17. Medical coding software. Collective Thinking. URL: https://www.collective-thinking.com/en/medical-coding-software/ [accessed 2022-03-14]

18. Facility coding: 3M™ 360 Encompass™ System for computer-assisted coding. 3M Health Information Systems. URL: https://www.3m.com/3M/en_US/health-information-systems-us/improve-revenue-cycle/coding/facility/360-encompass-computer-assisted-coding/ [accessed 2022-03-14]
19. Sumex Suite: The Sumex Suite is an established invoice verification solution tailored to the needs of Swiss insurance companies. ELCA. URL: https://www.elca.ch/en/sumex-suite [accessed 2022-03-14]
20. Liu Y, Cheng H, Klopfer R, Gormley MR, Schaaf T. Effective convolutional attention network for multi-label clinical document classification. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021 Presented at: EMNLP '21; November 7-11, 2021; Punta Cana, Dominican Republic p. 5941-5953. [doi: 10.18653/v1/2021.emnlp-main.481]
21. Yuan Z, Chuanqi T, Songfang H. Code synonyms do matter: multiple synonyms matching network for automatic ICD coding. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2022 Presented at: ACL '22; May 22-27, 2022; Dublin, Ireland p. 808-814. [doi: 10.18653/v1/2022.acl-short.91]
22. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv 2013 Jan 16. [doi: 10.48550/arXiv.1301.3781]
23. Mikolov T, Sutskever I, CHen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. In: Proceedings of the 26th International Conference on Neural Information Processing Systems. 2013 Presented at: NIPS '13; December 5-10, 2013; Lake Tahoe, NV, USA p. 3111-3119.
24. Le H, Vial L, Grej J, Segonne V, Coavoux M, Lecoteux B, et al. Flaubert: unsupervised language model pre-training for French. In: Proceedings of the 12th Language Resources and Evaluation Conference. 2020 Presented at: LREC '20; May 11-16, 2020; Marseille, France p. 2479-2490. [doi: 10.48550/arXiv.1912.05372]
25. Joulin A, Grave E, Bojanowski P, Mikolov T. Bag of tricks for efficient text classification. arXiv 2016 Jul 6. [doi: 10.48550/arXiv.1607.01759]
26. Wolf T, Debut L, Shah V, Chaumond J, Delangue C, Moi A, et al. Transformers: state-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. 2020 Presented at: EMNLP '20; November 16-20, 2020; Virtual p. 38-45. [doi: 10.18653/v1/2020.emnlp-demos.6]
27. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002 Jun 01;16:321-357. [doi: 10.1613/jair.953]

Abbreviations

AI: artificial intelligence
FlauBERT: French-Language Understanding via Bidirectional Encoder Representations from Transformers
ICD-10: International Statistical Classification of Diseases and Related Health Problems, 10th Revision
ML: machine learning
NLP: natural language processing
OOV: out of vocabulary
TF-IDF: term frequency-inverse document frequency

Edited by T Hao; submitted 21.03.22; peer-reviewed by S Puts, D Yu, K Rahmani; comments to author 19.06.22; revised version received 12.08.22; accepted 04.12.22; published 19.01.23.

Please cite as:
Xu HA, Maccari B, Guillain H, Herzen J, Agri F, Raisaro JL
An End-to-End Natural Language Processing Application for Prediction of Medical Case Coding Complexity: Algorithm Development and Validation
JMIR Med Inform 2023;11:e38150
URL: https://medinform.jmir.org/2023/1/e38150
doi:10.2196/38150
PMID:36656627

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org